Hilary Putnam

Razón, verdad e historia



tecnos

Hilary Putnam
Razón, verdad e historia

SERIE DE FILOSOFIA Y ENSAYO
Dirigida por Manuel Garrido

Alston, W. P.; Edwards, P.; Malcom, N.; Nelson, J. O., y Prior, A.: *Los orígenes de la filosofía analítica. Moore, Russell, Wittgenstein.*
Austin, J. L.: *Sentido y percepción.*
Boden, M. A.: *Inteligencia artificial y hombre natural.*
Bottomore, T.; Harris, L.; Kiernan, V. G.; Miliband, R.; con la colaboración de Kolakowski, L.: *Diccionario del pensamiento marxista.*
Brown, H. I.: *La nueva filosofía de la ciencia* (2.ª ed.)
Bunge, M.: *El problema mente-cerebro* (2.ª ed.).
Couturat, L.: *El álgebra de la lógica.*
Chisholm, R. M.: *Teoría del conocimiento.*
Dampier, W. C.: *Historia de la ciencia y sus relaciones con la filosofía y la religión.*
Díaz, E.: *Revisión de Unamuno.* Análisis crítico de su pensamiento político.
Doop, J.: *Nociones de lógica formal.*
Eccles, J. C.: *La psique humana.*
Edelman, B.: *La práctica ideológica del Derecho.*
Fann, K. T.: *El concepto de filosofía en Wittgenstein.*
Ferrater Mora, J., y otros: *Filosofía y ciencia en el pensamiento español contemporáneo (1960-1970).*
Feyerabend, P.: *Tratado contra el método.*
García Suárez, A.: *La lógica de la experiencia.*
Garrido, M.: *Lógica simbólica.*
Gómez García, P.: *La antropología estructural de Claude Lévi-Strauss.*
Gurméndez, C.: *Ser para no ser.* Ensayo de una dialéctica subjetiva.
Habermas, J.: *La lógica de las ciencias sociales.*
Habermas, J.: *Teoría y praxis.*
Hierro, J. S.-P.: *Problemas del análisis del lenguaje moral.*
Hintikka, J.: *Lógica, juegos de lenguaje e información.*
Hintikka, J.: *Saber y crecer.* Una introducción a la lógica de las dos nociones.
Kuhn, T. S.: *Segundos pensamientos sobre paradigmas.*
Lakatos, I., y otros: *Historia de la ciencia y sus reconstrucciones racionales* (2.ª ed.).
Lindsay, P. H., y Norman, D. A.: *Introducción a la psicología cognitiva* (2.ª ed.).
Lorenzen, P.: *Metamatemática.*
Lorenzo, J. de: *El método axiomático y sus creencias.*
Lorenzo, J. de: *La filosofía de la matemática de Jules Henri Poincaré.*
Martín Santos, L., y otros: *Ensayos de filosofía de la ciencia.*
Mates, B.: *Lógica matemática elemental.*
McCarthy, Th.; *La teoría crítica de Jürgen Habermas.*
París, C.: *Hombre y naturaleza.*
Popper, K. R.: *Búsqueda sin término.* Una autobiografía intelectual.
Popper, K. R.: *Realismo y el objetivo de la ciencia.* Post Scriptum a la Lógica de la investigación científica, vol. I.
Popper, K. R.: *El universo abierto.* Post Scriptum a la Lógica de la investigación científica, vol. II.
Popper, K. R.: *Teoría científica y el cisma en física.* Post Scriptum a la Lógica de la investigación científica, vol. III.
Prior, A. N.: *Historia de la lógica.*
Putnam, H.: *Razón, verdad e historia.*
Quine, W. V.: *La relatividad ontológica y otros ensayos.*
Quintanilla, M. A.: *Idealismo y filosofía de la ciencia.*
Rama, C. M.: *Teoría de la historia.* Introducción a los estudios históricos (3.ª ed.).
Rescher, N.: *La primacía en la práctica.*
Rivadulla, A.: *Filosofía actual de la ciencia.*
Robinet, A.: *Mitología, filosofía y cibernética.*
Rodríguez Paniagua, J. M.ª: *¿Derecho natural o axiología jurídica?*
Rodríguez Paniagua, J. M.ª: *Marx y el problema de la ideología.*
Sahakian, W. S.: *Historia y sistemas de la psicología.*
Smart, J. J. C., y Williams, B.: *Utilitarismo: pro y contra.*
Sotelo, I.: *Sartre y la razón dialéctica.*
Strawson, P. F.: *Ensayos lógico-lingüísticos.*
Vargas Machuca, R.: *El poder moral de la razón.* La filosofía de Gramsci.
Veldman, D. J.: *Programación de computadoras en ciencias de la conducta.*
Villacañas, J. L.: *Racionalidad crítica. Introducción a la filosofía de Kant.*
Wellman, C.: *Morales y éticas.*

RAZON, VERDAD E HISTORIA

HILARY PUTNAM

RAZON, VERDAD E HISTORIA

tecnos

Los derechos para la versión castellana
de la obra *Reason, Truth and History*
© Cambridge University Press, 1981
son propiedad de Editorial Tecnos, S.A.

Traducción:
José Miguel Esteban Cloquell

Impresión de cubierta:
Gráficas Molina

Josefa Valcárcel, 27 - 28027 Madrid
ISBN: 84-309-1577-X
Depósito Legal: M-21615-1988

Printed in Spain. Impreso en España por Azalso. Tracia, 17. Madrid.

A Ruth Anna

INDICE

PREFACIO

El propósito que anima la presente obra es acabar con la presión asfixiante que unas cuantas dicotomías parecen ejercer tanto sobre el pensamiento de los filósofos como sobre el de los legos. La principal de éstas es la dicotomía entre las concepciones objetivas y subjetivas de la verdad y de la razón.

El fenómeno que estoy considerando es el siguiente: una vez que una dicotomía como la existente entre lo «objetivo» y lo «subjetivo» se convierte en una dicotomía aceptada, y aceptada no como mero par de categorías, sino como una caracterización de tipos de concepciones y estilos de pensamiento, los pensadores comienzan a ver los términos de la dicotomía casi como etiquetas ideológicas. Muchos filósofos (quizá la mayoría) sostienen hoy alguna versión de la teoría de la verdad-copia, concepción de acuerdo con la cual un enunciado es verdadero sólo en el caso de que se «corresponda con los hechos (independientes de la mente)»; los filósofos de tal facción consideran que la única alternativa a ésta es negar la objetividad de la verdad y capitular con la idea de que todos los esquemas de pensamiento y todos los puntos de vista son desesperadamente subjetivos. Como es inevitable, una audaz minoría (Kuhn, por lo menos en algunos momentos, y algunos filósofos continentales tan distinguidos como Foucault) se alinea bajo la etiqueta opuesta. Estos últimos *están de acuerdo* en que la alternativa a la concepción ingenua de la verdad-copia es considerar subjetivos a los sistemas de pensamiento, a las ideologías, e incluso (en el caso de Kuhn y Feyerabend) a las teorías científicas, y pasan a *proponer* con vigor una perspectiva relativista y subjetiva.

No es de por sí necesariamente *nocivo* que la disputa filosófica asuma en parte el carácter de disputa ideológica: hasta en las ciencias más exactas, las nuevas ideas son reivindicadas y atacadas con vigor partisano. Incluso en política, la polarización y el fervor ideológico son a veces necesarios para darle seriedad moral a un asunto. Pero con el tiempo, y tanto en filosofía como en política, las nuevas ideas envejecen; lo que una vez fue un desafío se convierte en algo predecible y aburrido, y lo que una vez sirvió para centrar la atención allí donde debía centrarse, más tarde impide que la discusión tenga en cuenta nuevas alternativas. Esto ya ha ocurrido en el debate entre la teoría de la verdad-correspondencia y la perspectiva subjetivista. En los tres primeros capítulos de este libro intentaré exponer una concepción de la verdad que unifique los componentes objetivos y subje-

tivos. Esta concepción se retrotrae, al menos en su espíritu, a las ideas de Inmanuel Kant; y afirma que podemos rechazar la concepción ingenua de la verdad-copia sin tener que mantener que todo es cuestión de *Zeitgeist*, o de cambios gestálticos, o de ideología.

La concepción que voy a defender juzga, dicho sin rodeos, que hay una relación sumamente estrecha entre las nociones de *verdad* y de *racionalidad*; expresémoslo más directamente si cabe: que el único criterio para decidir lo que constituye un hecho es lo que es *racional* aceptar. (Con esto quiero afirmar algo completamente literal y sin excepción, de forma que si admitimos como posible el que sea racional aceptar que una pintura es bella, entonces es posible que sea un hecho el que la pintura es bella.) Mi concepción admite *hechos de valor*. Pero la relación entre la aceptabilidad racional y la verdad se da entre dos nociones distintas. Un enunciado puede ser racionalmente aceptable *en un tiempo* y no ser *verdadero*. En este trabajo trataré de preservar esta intuición realista.

Sin embargo, no creo que la racionalidad se defina mediante un conjunto de «cánones» o «principios» invariables; los principios metodológicos están relacionados con nuestra visión del mundo, incluyendo la visión que tenemos de nosotros mismos como una parte del mundo, y varían con el tiempo. De modo que *estoy de acuerdo* con los filósofos subjetivistas en que no hay ningún *organon* fijo y ahistórico que defina lo que es racional. Pero a partir del hecho de la evolución histórica de nuestras concepciones de la razón no concluyo que la razón pueda ser (o evolucionar hacia) algo, ni caigo en una fantástica mezcla de relativismo cultural y «estructuralismo», como hacen los filósofos franceses. Considero que la dicotomía: cánones de racionalidad ahistóricos e invariantes o relativismo cultural, está anticuada.

Otro rasgo de mi concepción es que no confina la racionalidad en el laboratorio científico, ni la considera fundamentalmente diferente dentro y fuera de éste. La idea contraria me parece una resaca del positivismo; de la idea de que el mundo científico está configurado en *cierta manera* por *«sense-data»* y de la idea de que los términos de las ciencias de laboratorio están definidos operacionalmente. No dedicaré mucho espacio a la crítica de las filosofías de la ciencia operacionalista y positivista, pues ya han sido criticadas exhaustivamente en la última veintena de años. Pero la idea empirista de que los *«sense data»* constituyen una especie de «planta baja» al menos para una parte de nuestro conocimiento, será reexaminada a la luz de lo que tenemos que decir a propósito de la verdad y de la racionalidad (capítulo 3).

En resumidas cuentas, presentaré una perspectiva según la cual la mente no «copia» simplemente un mundo que sólo admita la descripción de La Teoría Verdadera. Pero, desde mi punto de vista, la mente

no construye el mundo (ni siquiera en cuanto que estando sujeta a la constricción impuesta por «cánones metodológicos» y «*sense-data*» independientes de la mente). Y si es que nos vemos obligados a utilizar lenguaje metafórico, dejemos que la metáfora sea ésta: la mente y el mundo construyen conjuntamente la mente y el mundo (o, haciendo la metáfora todavía más hegeliana, el Universo construye el Universo —desempeñando nuestras mentes (colectivamente) un especial papel en la construcción).

Una de las finalidades de mi estudio acerca de la racionalidad es ésta: tratar de mostrar que nuestra noción de racionalidad es, en el fondo, solamente una parte de nuestra concepción del florecimiento humano, es decir, de nuestra idea de lo bueno. En el fondo, la verdad depende de lo que recientemente se ha denominado «valores» (capítulo 6). Y lo que afirmamos anteriormente con respecto a la racionalidad y a la historia, se aplica también al valor y a la historia; no hay ningún conjunto dado de «principios morales», ahistórico, que defina de una vez por todas en qué consiste el florecimiento humano; pero esto no significa que todo sea meramente cultural y relativo; ya que el presente estado de la teoría de la verdad —la dicotomía entre teorías de la verdad-copia y las formas subjetivas de considerar la verdad— es al menos parcialmente responsable (desde mi punto de vista) de la célebre dicotomía hecho/valor, tan sólo pasando a un nivel más profundo y corrigiendo nuestras concepciones de la verdad y de la racionalidad podremos ir más allá de esta dicotomía (dicotomía que, tal y como se entiende convencionalmente, nos compromete con algún tipo de relativismo). Los puntos de vista al uso sobre la verdad están enajenados; causan que se pierda una u otra parte de uno mismo y del mundo, y que se conciba al mundo consistiendo simplemente en partículas elementales desplazándose caóticamente en el vacío (la visión fisicalista, que considera que la descripción científica converge hacia La Teoría Verdadera) o en «*sense data* actuales o posibles» (la teoría empirista mas rancia) o a negar que haya un mundo como algo opuesto a un puñado de historias que fabricamos por diversas razones (principalmente inconscientes). Mi propósito en esta obra es esbozar las ideas directrices de una perspectiva no-enajenada.

Mi conferencia Herbert Spencer, «*Philosophers and Human Understanding*» (dada en la Universidad de Oxford en 1979) coincide en parte con el presente texto, y procede, como el artículo «"Si Dieu est mort, alors tout est permis"... (reflexions sur la philosophie du langage)», (*Critique*, 1980), del desarrollo de este mismo libro.

Una beca de investigación de la *National Science Foundation* [1] me

[1] Del estudio «The Appraisal of Scientific Theories: Historical versus Formal and Methodological Approaches» n.° SOC78-09276.

ayudó en una investigación relacionada con este libro, realizada durante los años 1978-80. Agradezco profundamente esta ayuda.

Thomas Kuhn y Ruth Anna Putnam han examinado esbozos de este libro y me han ofrecido críticas capaces y sabios consejos. También me ha ayudado la crítica de muchos amigos, entre ellos Ned Block, David Helman y Justin Leiber, y la de mis estudiantes de mis diversas clases y seminarios en Harvard. Varios capítulos los expuse como conferencia en Lima en la primavera de 1980 (en un viaje que fue posible gracias a una beca de la Comisión Fulbright) y el capítulo 2 fue finalizado durante mi estancia en esta ciudad. En este período me beneficié de las discusiones con Leopoldo Chiappo, Alberto Cordero Lecca, Henriques Fernández, Francisco Miro Quesada y Jorge Secada. Expuse todo el libro (en una primera versión) en forma de conferencias en la Universidad de Frankfurt, en el verano de 1980, y estoy en deuda con los colegas de esta Universidad (especialmente a Wilhem Essler y a Rainer Trapp), con mi muy estimulante grupo de alumnos y con mis demás amigos alemanes (especialmente Dieter Henrich, Manon Fassbinder y Wolfgang Stegmüller) por sus alentadoras e incitantes discusiones.

Todos mis colegas del Departamento de Filosofía de Harvard merecen ser reconocidos con un agradecimiento individual. En estos últimos años Nelson Goodman y yo hemos detectado una convergencia en nuestros puntos de vista y, pese a que escribí el primer borrador de este escrito sin haber tenido la oportunidad de consultar su libro *Ways of Worldmaking*, leer y discutir estos problemas con él ha tenido para mí un gran valor en varias fases.

También estoy en deuda con Jeremy Minott por su aliento y por mi confianza en su capacidad como editor.

1. CEREBROS EN UNA CUBETA

Una hormiga se arrastra lentamente sobre la arena. Conforme avanza va trazando en ésta una línea. Por puro azar, la línea se desvía y vuelve sobre sí misma, de tal forma que acaba pareciendo una reconocible caricatura de Winston Churchill. ¿Ha trazado la hormiga un retrato de Winston Churchill, un dibujo que *representa* a Churchill?

La mayoría de la gente, tras reflexionar un poco, contestaría que no. Después de todo, la hormiga nunca ha visto a Churchill, ni siquiera un retrato suyo, ni tampoco tenía intención de representarlo. Simplemente trazó una línea (y ni siquiera este acto fue intencional), línea que nosotros podemos ver como un retrato de Winston Churchill.

Podemos expresar esto afirmando que la línea no representa [1] por sí misma. La semejanza (de una especie muy complicada) con las facciones de Winston Churchill no es condición suficiente para que algo represente o se refiera a Churchill. Tampoco es condición necesaria: en nuestra comunidad, la forma impresa «Winston Churchill», las palabras «Winston Churchill», en tanto que pronunciadas, y muchas otras cosas, se usan para representar a Churchill (aunque no pictóricamente), si bien no tienen el tipo de semejanza con Churchill que sí tiene un retrato —o incluso un dibujo esquemático. Si la semejanza no es condición necesaria ni suficiente para que alguna cosa represente a otra, ¿cómo demonios puede una cosa representar (o estar en un lugar de, etc.) otra diferente?

La respuesta puede parecer fácil. Supongamos que la hormiga ha visto a Winston Churchill, y supongamos que tiene la inteligencia y la habilidad suficientes para dibujar un retrato suyo. Supongamos que ha elaborado la caricatura *intencionalmente*. Entonces la línea habría representado a Churchill.

Por otra parte, supongamos que la línea tiene la forma WINSTON

[1] En este libro, los términos «representación» y «referencia» aludirán siempre a la relación que se da entre una palabra (u otra clase de signo, símbolo o representación) y algo que existe efectivamente (esto es, no precisamente un «objeto del pensamiento»). Hay un sentido de «refiere» según el cual puedo «referirme» incluso a lo que no existe, mas no lo emplearé aquí. Una palabra más antigua para denominar lo que yo llamo «representación» o «referencia» es *denotación*. En segundo lugar, seguiré la costumbre de los lógicos modernos y usaré «existe» para significar «existe en el pasado, presente o futuro». Así pues, Winston Churchill «existe» y podemos «referirnos» o «representar» a Winston Churchill, aun cuando ya no esté vivo.

CHURCHILL, y que este hecho es un mero accidente (pasando por alto que es bastante improbable). Entonces los «caracteres impresos» WINSTON CHURCHILL no habrían representado a Winston Churchill, a pesar de que sí lo hacen cuando aparecen hoy en casi todos los libros.

De forma que puede antojársenos que lo que se necesita para la representación, o lo que se necesita principalmente para la representación, es la *intención.*

Pero para tener la intención de que *algo,* siquiera el lenguaje privado (incluso las palabras «Winston Churchill» repetidas mentalmente y no oídas) *represente* a Churchill, debo ser capaz de *pensar en* Churchill, para empezar. Si las líneas en la arena, los ruidos, etc., no pueden representar nada «en sí mismos», entonces ¿cómo es que pueden hacerlo las formas del pensamiento? ¿Cómo puede el pensamiento alcanzar y «aprehender» lo que es externo?

Algunos filósofos han dado un salto desde estas reflexiones hasta lo que ellos consideran como una prueba de *la naturaleza esencialmente no-física* de la mente. El argumento es simple; lo que dijimos acerca de la curva de la hormiga también se aplica a cualquier objeto físico. Ningún objeto físico tiene por sí mismo la capacidad de referirse a una cosa más bien que a otra; no obstante, es obvio que los *pensamientos de la mente* sí lo logran. De modo que los pensamientos (y por ende, la mente) poseen una naturaleza esencialmente distinta de la de los objetos físicos. Tienen la característica distintiva de la *intencionalidad* —pueden referirse a otras cosas; ningún objeto físico tiene «intencionalidad», salvo la intencionalidad que se deriva de su uso por parte de una mente. O eso se pretende. Pero esto es ir demasiado deprisa; postular misteriosos poderes mentales no resuelve nada. A pesar de todo el problema es real. ¿Cómo es posible la intencionalidad? ¿Cómo es posible la referencia?

TEORIAS MAGICAS DE LA REFERENCIA

Hemos visto que el «dibujo» trazado por la hormiga no tiene conexión necesaria con Winston Churchill. El mero hecho de que el dibujo mantenga cierta «semejanza» con Churchill no lo convierte ni en un retrato real ni en una representación de Churchill. Salvo que la hormiga sea una hormiga inteligente (y no es el caso) y sepa algo con respecto a Churchill (y tampoco es el caso), la curva que trazó no es un dibujo, ni tan siquiera una representación de algo. Ciertos pueblos primitivos creen que algunas representaciones (en particular los *nombres*) tienen una conexión necesaria con sus portadores; creen que saber el «verdadero nombre» de alguien o algo les otorga poder sobre ese alguien o algo. Este poder procede de una *conexión mágica*

entre el nombre y el portador del nombre; pero una vez que nos percatamos de que el nombre *sólo* tiene una conexión contextual, contingente y convencional con su portador, es difícil ver por qué el conocimiento del nombre ha de tener alguna significación mística.

Es importante darse cuenta de que a las *imágenes* mentales, y, en general, a las representaciones mentales, les ocurre lo mismo que a los dibujos físicos; la conexión que tienen las representaciones mentales con lo que representan no es más necesaria que la que tienen las representaciones físicas. La suposición contraria es un vestigio del pensamiento mágico.

El problema quizá se capte más fácilmente en el caso de las imágenes mentales (quizá el primer filósofo que captó la enorme importancia de este problema, pese a no ser realmente el primero en barajarlo, fue Wittgenstein). Supongamos que en alguna parte existe un planeta en el cual se han desarrollado seres humanos (o han sido depositados allí por extraños cosmonautas). Supongamos que esos humanos, si bien son como nosotros, nunca han visto un *árbol.* Supongamos que nunca se han imaginado un árbol (la única vida vegetal que existe en su planeta son los líquenes). Supongamos que cierto día, una nave que pasa por su planeta sin establecer contacto con ellos, arroja sobre éste el dibujo de un árbol. Imaginémosles devanándose los sesos ante el dibujo. ¿Qué demonios es esto? Se les ocurre toda clase de especulaciones: un edificio, un baldaquín, e incluso alguna especie de animal. Pero supongamos que ni siquiera se aproximan a saber de qué se trata.

Para *nosotros* la pintura es la representación de un árbol. Para aquellos humanos el dibujo únicamente representa un objeto extraño, de naturaleza y función desconocidas. Supongamos que, como resultado de ver el dibujo, uno de ellos tiene una imagen mental que es exactamente como mis imágenes mentales de los árboles. Su imagen mental no es la *representación de un árbol.* Sólo es la representación del extraño objeto (el que sea) que representa la misteriosa pintura.

Pese a esto, alguien podría argumentar que la imagen mental es *de hecho* la representación de un árbol, ya que, en primer lugar, el dibujo que provocó tal imagen mental era la representación de un árbol. Hay una cadena causal desde los árboles reales hasta la imagen mental, aun cuando esta sea muy extraña.

Sin embargo, podemos imaginarnos la ausencia de esta cadena. Supongamos que el «dibujo de un árbol» que la nave espacial arrojó no era en realidad el dibujo de un árbol, sino el resultado accidental del derrame de algunas pinturas. Aun cuando fuese exactamente igual al dibujo de un árbol, en realidad no sería el dibujo de un árbol en un grado mayor que la «caricatura» de la hormiga era un retrato de Churchill.

Podemos incluso imaginar que la nave espacial que arrojó el «dibujo» procedía de un planeta en el que no se sabía nada sobre los árboles. En tal caso, pese a que esos humanos tendrían imágenes cualitativamente idénticas a mi imagen de árbol, esas imágenes no representarían más a un árbol que a cualquier otra cosa arbitraria.

Lo mismo ocurre con las *palabras*. Un discurso impreso podría parecer una descripción perfecta de un árbol, pero si fueron los monos quienes lo produjeron golpeando fortuitamente las teclas de una máquina de escribir durante millones de años, entonces las palabras de ese discurso no se refieren a nada. Si alguien las memorizase y las repitiese mentalmente sin entenderlas, entonces cuando fuesen pensadas tampoco se referirían a nada.

Imaginemos que la persona que está repitiendo mentalmente estas palabras ha sido hipnotizada. Supongamos que tales palabras están en japonés, y que al hipnotizado se le ha dicho que entiende ese idioma. Supongamos que cuando piensa esas palabras experimenta algo así como un «sentimiento de comprensión». (Aunque si alguien irrumpiese en su flujo mental y le preguntase *qué significan* las palabras que está pensando, descubriría que no podría decirlo.) Quizá la ilusión sea tan perfecta que incluso podrá engañar a un telépata japonés. Pero si no es capaz de emplear las palabras en los contextos apropiados ni de responder preguntas con respecto a lo que «piensa», entonces no las entendió.

Combinando estas historias de ciencia-ficción, podemos idear un caso en el que alguien piensa ciertas palabras que constituyen de hecho la descripción de un árbol en algún lenguaje, y simultáneamente tiene unas imágenes mentales apropiadas, pero *ni* comprende las palabras *ni* sabe lo que es un árbol. Incluso podemos imaginar que las imágenes mentales fueron provocadas por un derrame de pintura (aunque la persona ha sido hipnotizada e inducida a pensar que son imágenes de algo apropiado a su pensamiento —sólo que, si se le pregunta de qué son imágenes, no podría responder). Y podemos imaginar que ni el hipnotizador ni el hipnotizado han oído hablar del lenguaje en el que este último está pensando— quizá sea una mera coincidencia el que estas «oraciones sin sentido», tal y como las considera el hipnotizador, sean la descripción de un árbol en idioma japonés. En resumidas cuentas, cualquier cosa que pase ante su mente puede ser cualitativamente idéntica a lo que estaba pasando por la mente de un hablante japonés que pensaba *verdaderamente* en árboles— pero ninguna de ellas se referiría a árboles.

Todo esto es realmente imposible, por supuesto, del mismo modo que es realmente imposible que los monos mecanografíen por casualidad una copia de *Hamlet*. Y esto es afirmar que las posibilidades en contra son demasiado altas como para que este suceso realmente ocurra. A pesar de todo, no es lógicamente imposible, ni siquiera físi-

camente imposible. *Podría* suceder (es compatible con las leyes de la física, y quizá también con las condiciones actuales del universo, si hubiese seres inteligentes sobre otros planetas). Y si sucediese, sería una sorprendente demostración de una importante verdad conceptual: ni siquiera un amplio y complejo sistema de representaciones verbales y visuales tiene una conexión *intrínseca,* mágica, dada de una vez por todas, con lo que representa; una conexión independiente del modo en que fue causada y de lo que constituyen las disposiciones del sujeto hablante o pensante. Y esto es cierto tanto si el sistema de representaciones (palabras e imágenes, en nuestro ejemplo) esta implementado físicamente —las palabras son palabras escritas o habladas y los dibujos son dibujos físicos— o tan sólo concebido mentalmente. Ni las palabras del pensamiento ni las imágenes mentales representan *intrínsecamente* aquello acerca de lo que tratan.

EL CASO DE LOS CEREBROS EN UNA CUBETA

He aquí una posibilidad de ciencia-ficción discutida por los filósofos: imaginemos que un ser humano (el lector puede imaginar que es él quien sufre el percance) ha sido sometido a una operación por un diabólico científico. El cerebro de tal persona (su cerebro, querido lector) ha sido extraído del cuerpo y colocado en una cubeta de nutrientes que lo mantienen vivo. Las terminaciones nerviosas han sido conectadas a una computadora supercientífica que provoca en esa persona la ilusión de que todo es perfectamente normal. Parece haber gente, objetos, cielo, etc.; pero en realidad todo lo que la persona (usted) está experimentando es resultado de impulsos electrónicos que se desplazan desde la computadora hasta las terminaciones nerviosas. La computadora es tan ingeniosa que si la persona intenta alzar su mano, el «*feedback*» que procede de la computadora le provocará que «vea» y «sienta» que su mano está alzándose. Por otra parte, mediante una simple modificación del programa, el diabólico científico puede provocar que la víctima «experimente» (o alucine) cualquier situación o entorno que él desee. También puede borrar la memoria de funcionamiento del cerebro, de modo que la víctima crea que siempre ha estado en ese entorno. La víctima puede creer incluso que está sentado, leyendo estas mismas palabras acerca de la suposición, divertida aunque bastante absurda, de que hay un diabólico científico que extrae cerebros de los cuerpos y los coloca en una cubeta de nutrientes que los mantiene vivos. Las terminaciones nerviosas se suponen conectadas a una computadora supercientífica que provoca en la persona la ilusión de ...

Cuando se menciona esta especie de posibilidad en una clase de Teoría del Conocimiento, el propósito no es otro que suscitar de un

modo moderno el clásico problema del escepticismo con respecto al mundo externo. *(¿Cómo podría usted saber que no se halla en esa situación?)* Pero esta situación es también un útil recurso para suscitar cuestiones en torno a la relación mente-mundo.

En lugar de imaginar un solo cerebro en una cubeta, podemos imaginar que los seres humanos (quizá todos los seres sintientes) son cerebros en una cubeta (o sistemas nerviosos en una cubeta, en el caso de algunos seres que sólo poseen un sistema nervioso mínimo, pero que ya cuentan como sintientes). Por supuesto, el diabólico científico tendría que estar fuera —¿o querría estarlo? Quizá no exista ningún diabólico científico, quizá (aunque esto es absurdo) el mundo consista en una maquinaria automática que está al cuidado de una cubeta repleta de cerebros y sistemas nerviosos.

Supongamos esta vez que la maquinaria automática está programada para ofrecernos a todos una alucinación *colectiva,* en lugar de unas cuantas alucinaciones separadas y sin relación. De forma que cuando me parece estar hablando con usted, a usted le parece estar oyendo mis palabras. Mis palabras no llegan realmente a sus oídos, por supuesto —porque usted no tiene oídos (reales), ni yo tengo boca o lengua reales. Pero cuando emito mis palabras, lo que ocurre en realidad es que los impulsos aferentes se desplazan desde mi cerebro hasta el ordenador, el cual a su vez provoca que yo «oiga» mi propia voz profiriendo esas palabras y «sienta» el movimiento de mi lengua, y que usted «oiga» mis palabras, y me «vea» hablando, etc. En este caso, nos comunicamos realmente, hasta cierto punto. Yo no estoy equivocado con respecto a su existencia real (sólo lo estoy con respecto a la existencia de su cuerpo y del «mundo externo», aparte de los cerebros). En cierta medida, tampoco importa que «el mundo entero» sea una alucinación colectiva; después de todo, cuando me dirijo a usted, usted oye realmente mis palabras, si bien el mecanismo no es el que suponemos. (Si fuéramos dos amantes haciendo el amor y no dos personas manteniendo una conversación, la insinuación de que únicamente somos dos cerebros en una cubeta podría ser molesta, desde luego.)

Deseo formular ahora una pregunta que parecerá obvia y bastante estúpida (al menos a algunos, incluyendo a algunos filósofos sumamente sofisticados), pero que tal vez nos sumerja con cierta rapidez en auténticas profundidades filosóficas. Supongamos que toda esta historia fuera realmente verdadera. Si fuéramos cerebros en una cubeta, ¿podríamos *decir o pensar* que lo somos?

Voy a argumentar en favor de la respuesta «no, no podríamos». En realidad, voy a argüir que la suposición de que realmente somos cerebros en una cubeta, pese a no violar ley física alguna y a ser perfectamente consistente con todas nuestras experiencias, no puede ser

verdadera. Y no *puede ser verdadera* porque, en cierto modo, se autorrefuta.

El argumento que seguidamente expondré es bastante inusual, y me llevó varios años convencerme de que era verdaderamente correcto. Pero es un argumento correcto. Lo que le da una apariencia tan extraña es su conexión con algunas de las más profundas cuestiones de la filosofía (se me ocurrió por primera vez cuando estaba estudiando un teorema de la lógica moderna, el teorema de Skolem-Löwenheim, y vi de repente una conexión entre este teorema y algunos argumentos de las *Investigaciones Filosóficas* de Wittgenstein).

Un «supuesto que se autorrefuta» es aquel cuya verdad implica su propia falsedad. Por ejemplo, consideremos la tesis de que *todos los enunciados generales son falsos.* Este es un enunciado general. De forma que si es verdadero, debe ser falso. Por lo tanto es falso. En ciertas ocasiones decimos que una tesis se «autorrefuta» si *la misma suposición de que la tesis es tomada en cuenta o enunciada* implica ya su falsedad. Por ejemplo, «no existo» se autorrefuta si *soy yo* (para cualquier «*yo*») quien lo pienso. De modo que uno puede estar seguro de que existe con sólo pensar en ello (como Descartes argumentó).

Demostraré que la suposición de que somos cerebros en una cubeta posee precisamente esta propiedad. Si podemos considerar su verdad o su falsedad, entonces no es verdadera (lo demostraré). Por lo tanto no es verdadera.

Antes de ofrecer el argumento, permítanme considerar el motivo por el que parece tan extraño que éste pueda siquiera ofrecerse (al menos para los filósofos que subscriben una concepción de la verdad-copia). Concedamos que es compatible con las leyes físicas que haya un «mundo posible» en el que todos los seres sintientes sean cerebros en una cubeta. (El discurso sobre este «mundo posible» suena como si hubiese un *lugar* donde cualquier suposición absurda fuese verdadera, y éste es el motivo por el que puede ser filosóficamente desorientador.) Los humanos de ese mundo posible tienen exactamente las mismas experiencias que tenemos *nosotros.* También piensan igual que nosotros (al menos por su mente pasan palabras, imágenes, formas de pensamiento). Aún así, estoy afirmando que podemos ofrecer un argumento que demuestre que no somos cerebros en una cubeta. ¿Cómo puede haberlo? ¿Por qué no podrían ofrecer tal argumento las personas que, en tal mundo posible, *son* realmente cerebros en una cubeta?

La respuesta será (básicamente) ésta: aunque esas personas pueden pensar y «decir» cualquier palabra que nosotros pensemos o digamos, no pueden «referirse» a lo que nosotros nos referimos. En particular, no pueden decir o pensar que son cerebros en una cubeta *(incluso pensando «somos cerebros en una cubeta»).*

EL TEST DE TURING

Supongamos que alguien ha logrado inventar una computadora que pueda mantener realmente una conversación con un sujeto (o con tantas personas como podría hacerlo una persona inteligente) ¿Cómo podríamos decidir si la computadora es «consciente» o no?

El lógico británico A. Turing propuso el siguiente test[2]: permitamos que un sujeto mantenga una conversación con la computadora y otra conversación con una persona a quien no conoce. Si no puede decidir quién es la computadora y quién el ser humano, entonces (suponiendo que el test se repite en número suficiente de veces con distintos interlocutores), la computadora es consciente. En resumen, una máquina de computar es consciente si puede pasar el «Test de Turing» (las conversaciones no se mantienen cara a cara, por supuesto: el interlocutor no puede conocer la apariencia visual de sus dos contertulios; tampoco debe utilizarse la voz, ya que la voz mecánica podría sonar con un matiz distinto al de la voz humana). Imaginemos, más bien, que todas las conversaciones se mantienen con el soporte de una máquina de escribir eléctrica. El interlocutor mecanografía sus enunciados, preguntas, etc., y ambos contertulios —la máquina y el ser humano— responden por medio de un teclado eléctrico. (La máquina puede además *mentir*— si se le pregunta «¿Eres una máquina?», podría replicar «No, soy un asistente del laboratorio».)

Varios autores (de ningún modo hostiles a la idea de que una máquina pueda ser consciente) han criticado la idea de que este test sea realmente definitivo para la determinación de la conciencia. Pero nuestro tema no es éste. Deseo emplear la idea general del Test de Turing, la idea general de un *test dialógico de competencia,* para un propósito distinto: explorar la noción de *referencia.*

Imaginemos una situación en la que el problema no consiste en determinar si el contertulio es realmente una persona o una máquina, sino más bien determinar si usa referencialmente las palabras tal y como nosotros lo hacemos. El test más obvio consiste, de nuevo, en matener una conversación y, si no surgen problemas, si el contertulio «aprueba» (en el sentido de que no puede ser distinguido de alguien de quien sabemos de antemano que habla el mismo lenguaje, se refiere a las mismas clases usuales de objetos, etc.), concluir que se refiere a los objetos tal y como nosotros lo hacemos. Cuando el propósito del Test de Turing sea precisamente el descrito, es decir, determinar la referencia compartida *(shared)* lo llamaré *Test de Turing para la Referencia.* Y del mismo modo que los filósofos han debatido la cues-

---

[2] A.M. TURING, «Computing Machines and Intelligence», *Mind,* 1950, reimpreso en A. R. Anderson (ed.), *Minds and Machines.*

tión de si el original Test de Turing constituye un test *definitivo* para la determinación de la conciencia, esto es, la cuestión de si una máquina que «apruebe» el test, no en una sola ocasión, sino regularmente, es *necesariamente* consciente, deseo discutir la cuestión de si el recién sugerido Test de Turing para la Referencia constituye un test definitivo para determinar la referencia compartida.

La respuesta resultará ser «NO». El Test de Turing para la Referencia no es definitivo. No hay duda de que es un test excelente en la práctica, pero no es lógicamente imposible (aunque sin duda es altamente improbable) que alguien pase el Test de Turing sin estar refiriéndose a nada. De aquí se sigue, como veremos, que es posible ampliar nuestra observación de que las palabras (y los textos completos y los discursos) no tienen una conexión necesaria con sus referentes. Aun cuando consideremos las reglas que deciden qué palabras pueden emitirse apropiadamente en ciertos contextos, en lugar de las palabras como tales —aun cuando consideremos *programas para el uso de las palabras,* según la jerga informática—, aún así, esas palabras no poseen una referencia determinada, a menos que esos programas se *refieran* por sí mismos *a algo extralingüístico.* Este será un paso crucial de cara a alcanzar la conclusión de que los cerebros en una cubeta no pueden referirse de ningún modo a nada externo (y, por consiguiente, no pueden decir *que* son cerebros en una cubeta).

Supongamos, por ejemplo, que me encuentro en la situación ideada por Turing (disputando el «juego de la imitación», según la terminología de Turing) y que mi contertulio es efectivamente una máquina. Supongamos que la máquina es capaz de ganar el juego («pasa» el test). Imaginemos que la máquina está programada para producir bellas respuestas en castellano en contestación a enunciados, preguntas, observaciones, etc., formulados en castellano, pero que no tiene órganos sensoriales (salvo la conexión con mi máquina de escribir eléctrica). (Turing no da por sentado que la posesión de otros órganos sensoriales o motores sea necesaria para la conciencia o la inteligencia, que yo sepa.) Supongamos que la máquina no sólo carece de ojos y oídos electrónicos, etc., sino que su programa, el programa que le permite llevar a cabo el juego de la imitación, también carece de dispositivos para la incorporación de *inputs* desde esos órganos, o para ejercer control sobre un cuerpo. ¿Qué tenemos que decir con respecto a esta máquina?

Me parece evidente que ni podemos ni debemos atribuir uso referencial a tal ingenio mecánico. Es cierto que la máquina puede platicar maravillosamente acerca del paisaje de Nueva Inglaterra, por ejemplo. Pero si tuviese enfrente una manzana, una montaña o una vaca, no podría reconocerlas.

Tenemos un artilugio para producir oraciones en respuesta a ora-

ciones. Pero estas oraciones no están conectadas con el mundo real. *Si acopläsemos* dos de estas máquinas y les dejásemos disputar el juego de la imitación entre sí, *continuarían «engañándose» una a otra eternamente, aun cuando el resto del mundo desapareciese.* No existen más razones para considerar que el discurso de la máquina acerca de manzanas se refiere a manzanas del mundo real, que para considerar que el «retrato» de la hormiga se refiere a Winston Churchill.

Lo que produce aquí la ilusión de referencia, significado, inteligencia, etc., es el hecho de que *tenemos* una convención de representación bajo la cual el discurso de la máquina se refiere a manzanas, campanarios, Nueva Inglaterra, etc. De modo parecido, y por la misma razón, existe la *ilusión* de que la hormiga ha caricaturizado a Churchill. Sólo que *nosotros* somos capaces de percibir, manipular y mantener trato con manzanas y campos. Existen «reglas de entrada al lenguaje» que nos conducen desde las experiencias con manzanas a proferencias tales como «Veo una manzana» y «reglas de salida del lenguaje» que nos llevan desde decisiones expresadas en forma lingüística («Voy a comprar algunas manzanas») a acciones distintas de la acción de hablar. Como la máquina carece de reglas de entrada o de salida del lenguaje, no hay ninguna razón para considerar el discurso de la máquina (o de ambas máquinas, en el caso que ideamos de dos máquinas jugando entre sí el juego de la imitación), como algo más que un juego sintáctico. Un juego sintáctico que se *parece* a un discurso inteligente, a buen seguro. Pero en la misma medida (y no en mayor) que la curva de la hormiga se parece a una mordaz caricatura.

En este último caso, podríamos haber argumentado que la hormiga habría dibujado la misma curva aunque Winston Churchill jamás hubiese existido. Pero no podemos elaborar un argumento que sea completamente paralelo para el caso de la máquina; si las manzanas, los árboles y los campanarios no hubieran existido, entonces, presumiblemente, los programadores no habrían dado a luz el mismo programa. A pesar de que la máquina no *percibe* manzanas, campos o campanarios, sus creadores-diseñadores sí lo hacen. Hay *cierta* conexión causal entre la máquina y las manzanas del mundo real, etc., por medio de la experiencia perceptiva y el conocimiento de sus creadores-diseñadores. Pero esta débil conexión difícilmente puede bastar para la referencia. No sólo es lógicamente posible —pese a ser fantásticamente improbable— que la misma máquina hubiera existido sin existir manzanas, campos y campanarios, etc. Aun cuando todas estas cosas *cesaran* de existir, la máquina seguiría platicando felizmente. Por este motivo, es absolutamente imposible sostener que la máquina es capaz de referirse a las cosas.

El punto relevante para nuestra discusión es que el Test de Turing no permite excluir a una máquina programada para no hacer otra cosa *salvo disputar* el Juego de la Imitación, y está claro que una má-

quina con tales características no se refiere a las cosas en mayor medida en que lo hace un tocadiscos.

CEREBROS EN UNA CUBETA (DE NUEVO)

Permítasenos comparar los hipotéticos «cerebros en una cubeta» con las máquinas recién descritas. Obviamente, hay importantes diferencias. Los cerebros en una cubeta no tienen órganos sensoriales, aunque sí tienen elementos que funcionan como tales; es decir, hay terminaciones nerviosas aferentes e *inputs,* y estos *inputs* figuran en el programa de los cerebros en una cubeta en la misma medida en que figuran en el nuestro. Los cerebros en una cubeta son *cerebros;* por otra parte, son cerebros que *funcionan,* y lo hacen mediante las mismas reglas con las que funcionan los cerebros del mundo real. Por estas razones, parecería absurdo negarles conciencia o inteligencia. Pero el hecho de que sean conscientes o inteligentes no significa que sus palabras se refieran a lo que se refieren las nuestras. La cuestión que nos interesa es: cuando sus verbalizaciones contienen la palabra «árbol», ¿se refieren realmente a árboles? De forma más general: ¿acaso pueden referirse a objetos *externos?* (Como algo opuesto a los objetos que aparecen en la imagen producida por la maquinaria automática, por ejemplo.)

En orden a organizar nuestras ideas, permítasenos suponer que la máquina automática existe gracias a cierto tipo de azar o coincidencia cósmica (o quizá que siempre ha existido). En este hipotético mundo, se supone que la maquinaria automática no tiene un creador-diseñador inteligente. Como dijimos al comenzar este capítulo, podemos imaginar que todas las criaturas sensibles (aunque su sensibilidad sea mínima) se hallan en el interior de la cubeta.

Esta suposición no nos ayuda. Porque no hay conexión alguna entre la *palabra* «árbol» —tal como es utilizada por estos cerebros— y los árboles reales. Aunque no hubieran árboles reales, seguirían usando la palabra «árbol» como lo hacen, pensando los mismos pensamientos que piensan y teniendo exactamente las mismas imágenes que tienen. Sus imágenes, palabras, etc., son cualitativamente idénticas a las imágenes, las palabras, etc., que sí representan árboles en *nuestro* mundo; pero ya hemos visto (¡la hormiga otra vez!) que la semejanza cualitativa con algo que representa un objeto (Winston Churchill o un árbol) no hace que una cosa sea por sí misma una representación. En resumen, cuando los cerebros en una cubeta piensan «hay un árbol delante de mí» no están pensando en árboles reales, ya que no existe nada en virtud de lo cual su pensamiento «árbol» represente árboles reales.

Si esta conclusión parece precipitada, reflexionemos sobre lo si-

guiente: hemos visto que las palabras no se refieren necesariamente a árboles pese a estar ordenadas en una secuencia idéntica a un discurso que (si ocurriese en una de nuestras mentes) se *referiría* incuestionablemente a *los árboles* del mundo real. Tampoco «el programa», en el sentido de las reglas, prácticas y disposiciones de los cerebros a la conducta verbal, se refiere necesariamente a árboles, ni ocasiona la referencia a árboles a través de las conexiones que establece entre palabras y palabras, o entre señales de entrada *lingüística* y respuestas *lingüísticas*. Si estos cerebros representan, se refieren y piensan acerca de árboles (árboles reales, fuera de la cubeta), debe ser gracias al modo en que el programa conecta el sistema lingüístico con *inputs y outputs no-verbales*. Tales *inputs* existen efectivamente en el mundo de los cerebros en una cubeta (¡de nuevo aquellas terminaciones nerviosas aferentes y eferentes!) pero también veíamos que los «*sense-data*» producidos por la maquinaria automática no representan árboles (o algo externo) aun cuando se parezcan con exactitud a nuestras imágenes de los árboles. Así como una mancha de pintura podría parecer el dibujo de un árbol sin ser el dibujo de un árbol, veíamos que un «dato sensorial» podría ser cualitativamente idéntico a la «imagen de un árbol», sin ser la imagen de un árbol. En el caso de los cerebros en una cubeta, ¿cómo es que el hecho de que el lenguaje está conectado mediante el programa con *inputs* sensoriales que no representan intrínseca o extrínsecamente árboles (o cualquier otra cosa externa) puede posibilitar que todo el sistema de representaciones, el lenguaje-en-uso, represente o se refiera a árboles, o a cualquier otra cosa externa?

La respuesta es que no puede. Todo el sistema de *sense-data,* las señales motoras para las terminaciones eferentes y el pensamiento mediatizado verbal o conceptualmente y conectado mediante «reglas de entrada al lenguaje» con los *sense-data* (o con lo que sea) como *inputs* y mediante «reglas de salida del lenguaje» con las señales motoras como *outputs,* no tiene más conexiones con los *árboles* que las que la curva de la hormiga tenía con Winston Churchill. Una vez vemos que la *semejanza cualitativa* (ascendiéndola, si queremos, a identidad cualitativa) entre los pensamientos de los cerebros en una cubeta y los de alguien que exista en el mundo real no implica en modo alguno la mismidad de la referencia, no es difícil ver que no hay ninguna base para considerar que el cerebro en una cubeta se refiera a cosas externas.

LAS PREMISAS DEL ARGUMENTO

Ya he ofrecido el argumento prometido para demostrar que los cerebros en una cubeta no pueden pensar ni decir que son cerebros

en una cubeta. Queda sólo hacerlo explícito y examinar su estructura.

Según acabamos de afirmar, cuando el cerebro en una cubeta (en aquel mundo donde cada ser sintiente está y siempre estuvo dentro de una cubeta) piensa «Hay un árbol delante de mí», su pensamiento no se refiere a árboles reales. Según algunas teorías que seguidamente discutiremos, podría referirse a árboles que aparecen en la imagen, o a impulsos electrónicos que ocasionan experiencias de árboles, o a las características del programa que son responsables de esos impulsos electrónicos. Nuestras recientes afirmaciones no implican el rechazo de estas teorías, pues existe una estrecha conexión causal entre el uso de la palabra «árbol» en la cubeta-castellana y la presencia de árboles que aparecen en la imagen, la presencia de impulsos electrónicos de cierto tipo, y la presencia de ciertas características en el programa de la máquina. Según tales teorías, el cerebro *está en lo cierto, no se equivoca* al pensar «Hay un árbol delante de mí». Dada la referencia de «árbol» y de «delante de» en la cubeta-castellana, y suponiendo que una de estas teorías es correcta, entonces las condiciones de verdad de la oración «Hay un árbol delante de mí», cuando ocurre en una cubeta-castellana, son algo tan simple como el que haya un árbol aparente-en-la-imagen delante del mí en cuestión —en la imagen— o quizá que desde la maquinaria esté llegando el tipo de impulso electrónico que normalmente produce esta experiencia, o quizá que esté operando el dispositivo de la maquinaria que se supone que produce la experiencia de «árbol delante de mí». Y sin duda alguna, estas condiciones de verdad se satisfacen.

Por el mismo argumento, en la cubeta-castellana «cubeta» se refiere a cubetas-aparentes-en-la-imagen, o a algo relacionado (impulsos electrónicos o a características de programa), pero sin duda no se refiere a cubetas reales, ya que el uso de «cubeta» en la cubeta-castellana no tiene conexión causal con cubetas reales (aparte de la conexión que supone el hecho de que los cerebros en una cubeta no podrían usar la palabra «cubeta» si no estuvieran en presencia de una cubeta particular —la cubeta en que se hallan; pero en la cubeta-castellana esta conexión se obtiene entre el uso de toda palabra y esa cubeta particular: no es una conexión especial entre el uso de la palabra «cubeta» y las cubetas). De forma similar, en la cubeta-castellana, «fluido nutriente» se refiere un líquido-aparente-en-la-imagen o a algo relacionado (impulsos electrónicos o características del programa). Se sigue que si su «mundo posible» es el mundo real, y somos realmente cerebros en una cubeta, entonces lo que queremos decir con «Somos cerebros en una cubeta» es que *somos cerebros en una cubeta-aparente-en-la-imagen* o algo de esta índole (si es que queremos decir algo). Pero parte de la hipótesis de que somos cerebros en una cubeta

es que no somos cerebros en una cubeta-aparente-en-la-imagen (es decir, lo que estamos «alucinando» no es que somos cerebros en una cubeta). Así pues, si somos cerebros en una cubeta, entonces la oración «Somos cerebros en una cubeta» afirma algo falso (si es que afirma algo). En resumidas cuentas, si somos cerebros en una cubeta, entonces «Somos cerebros en una cubeta» es falso. Por lo tanto es (necesariamente) falso.

La suposición de que tal posibilidad tenga sentido surge de la combinación de dos errores: (1) tomar demasiado en serio la *posibilidad física;* y (2) manejar de forma inconsciente una teoría mágica de la referencia, es decir, una teoría en la que ciertas representaciones mentales se refieren necesariamente a ciertas cosas y clases de cosas externas.

Hay un «mundo físicamente posible» en el cual somos cerebros en una cubeta —¿qué significa esto, excepto que hay una descripción de tal estado de cosas que es compatible con las leyes de la física? Así como existe en nuestra cultura (desde el siglo XVIII) la tendencia a considerar la física como nuestra metafísica, es decir, a considerar las ciencias exactas como la descripción del «verdadero mobiliario último del universo» por tanto tiempo buscada, la consecuencia inmediata de esta tendencia es también cierta tendencia a considerar la «posibilidad física» como la mismísima piedra de toque de cómo podrían ser verdadera y realmente las cosas. La verdad es verdad física, la posibilidad, posibilidad física, y la necesidad, necesidad física, según esta opinión. Pero acabamos de ver, aunque por ahora sólo en el caso de un ejemplo muy artificial, que este punto de vista es erróneo. La existencia de un «mundo físicamente posible» en el cual somos cerebros en una cubeta (y siempre lo fuimos y siempre lo seremos) no significa que posible, verdadera y realmente podríamos *ser* cerebros en una cubeta. Lo que excluye esta posibilidad no es la física, sino la *filosofía.*

Algunos filósofos, anhelando hacer valer y al mismo tiempo minimizar las pretensiones de su profesión (típico estado de ánimo en la filosofía angloamericana del siglo XX) dirían: «Bravo. Ha mostrado que algunas cosas que parecen ser posibilidades físicas son en realidad imposibilidades *conceptuales.* ¿Qué tiene esto de sorprendente?»

Bien, no hay duda de que mi argumento puede describirse como un argumento conceptual. Pero cuando la actividad filosófica se describe como una búsqueda de verdades «conceptuales», suena a algo así como una *investigación acerca del significado de las palabras.* Y, no ha sido ésta nuestra tarea, desde luego.

Nuestra tarea ha consistido en examinar las *precondiciones* del *pensar acerca de algo, representar algo, referirse a algo,* etc. Y no hemos investigado estas precondiciones desentrañando el significado de es-

tas palabras y de estas frases (como podría hacer un lingüista, por ejemplo), sino *razonando a priori.* Y no en el antiguo sentido «absoluto» (ya que no pretendemos que las teorías mágicas de la referencia sean erróneas *a priori*), sino en el sentido de una investigación sobre lo que es *razonablemente* posible una vez *asumidas* ciertas premisas generales, o una vez *establecidas* ciertas suposiciones teóricas muy generales. Tal procedimiento no es «empírico», pero tampoco es completamente «a priori», aunque incorpora elementos de ambas formas de investigación. A pesar de ser falible y de depender de supuestos que podrían ser descritos como «empíricos» (por ejemplo, el supuesto de que la mente no tiene otro acceso a las cosas externas o a las propiedades que el que le proporcionan los sentidos), mi procedimiento se halla en estrecha relación con lo que Kant llamó «investigación trascendental», ya que consiste, repito, en una investigación sobre las *precondiciones* de la referencia y, por ende, del pensamiento —precondiciones que se encuentran incorporadas en la naturaleza de nuestras mentes, aunque no son (como Kant creía) completamente independientes de suposiciones empíricas.

Una de las premisas del argumento es obvia: las teorías mágicas de la referencia son erróneas, y no sólo erróneas con respecto a las representaciones físicas, sino también en lo que concierne a las mentales. La otra premisa establece la imposibilidad de referirnos a ciertos tipos de cosas, por ejemplo, *a árboles,* sin haber tenido interacción causal con ellas [3], o, en su caso, con otras cosas en cuyos términos puedan describirse las primeras. Pero ¿por qué hemos de aceptar estas premisas? Puesto que constituyen el marco general en el que opera mi argumentación, es el momento de examinarlas más detenidamente.

RAZONES PARA NEGAR QUE HAYAN CONEXIONES NECESARIAS ENTRE LAS REPRESENTACIONES Y SUS REFERENTES

Anteriormente apuntaba que algunos filósofos (el más célebre, Brentano) han adscrito una facultad a la mente, la «intencionalidad», que la capacita precisamente para *referirse* a las cosas. Evidentemente, descarto que esto constituya una solución. Pero ¿con qué derecho lo hago? ¿He ido demasiado deprisa?

Estos filósofos no afirmaban que fuera posible pensar en cosas

[3] Si los cerebros en una cubeta fueran a tener conexión causal con árboles en el *futuro,* por ejemplo, entonces quizá podrían referirse *ahora* a árboles mediante la descripción «Las cosas a las que me referiré como «árboles» en tal y cual ocasión futura». Pero, en nuestro caso, los cerebros en una cubeta no se libran *nunca* de la cubeta, y por *tanto* no entran *nunca* en conexión causal con árboles, etc.

externas o en propiedades sin usar ningún tipo de representación. Y aceptarían el argumento que ofrecí anteriormente y que comparaba los *sense data* con el «dibujo» de la hormiga (el argumento que se servía de la historia de ciencia ficción acerca del «dibujo» de un árbol producido por una mancha de pintura, que daba lugar a *sense data* cualitativamente similares a nuestras imágenes visuales de los árboles, pero que no estaban acompañados por ningún *concepto* de árbol), como muestra de que las *imágenes* no se refieren a las cosas de un modo necesario. Si existen representaciones mentales que se refieran necesariamente (a cosas externas), deben ser de la naturaleza de los *conceptos,* no de las imágenes. Pero ¿qué son los *conceptos?*

Cuando practicamos la introspección no percibimos «conceptos» fluyendo por nuestra mente así como así. Dondequiera y cuandoquiera que detengamos el flujo de pensamiento, lo que atrapamos son palabras, imágenes, sensaciones, sentimientos. Cuando digo en voz alta mis pensamientos, no los pienso por segunda vez. Oigo mis palabras como usted la oye. No hay duda de que cuando creo las palabras que profiero me siento en un estado «diferente» a cuando no las creo (pero a veces, cuando estoy nervioso, o ante una audiencia hostil, me parece estar mintiendo, pese a saber que estoy diciendo la verdad); también me siento en un estado distinto cuando profiero palabras que comprendo a cuando profiero palabras que no comprendo. Pero puedo imaginar sin dificultad a alguien que piense estas palabras (en el sentido de repetirlas mentalmente) y experimente el mismo sentimiento de comprensión, afirmación, etc., que yo tengo, pero percatándose (o siendo despertado por un hipnotizador) un minuto más tarde de que no entendía lo que acababa de pasar por su mente, y de que ni siquiera entendía el lenguaje en el que estaban esas palabras. No pretendo que esto sea muy probable, sino que es perfectamente imaginable. Y esto no muestra que los conceptos *son* palabras (o imágenes, sensaciones, etc.) sino que atribuir a alguien un concepto o un pensamiento es algo completamente distinto de atribuirle alguna «representación» mental, alguna entidad o evento susceptible de introspección. La razón decisiva para sostener que los conceptos no sean representaciones mentales que se refieran intrínsecamente a las cosas es que ni tan siquiera son representaciones mentales. Los conceptos son símbolos que se usan de cierto modo; los símbolos pueden ser públicos o privados, entidades mentales o físicas, pero aun cuando los símbolos sean «mentales» y «privados», el propio símbolo, con independencia de su uso, no es el concepto. Y los símbolos no se refieren de por sí intrínsecamente a ninguna cosa.

Podemos observar esto realizando un experimento mental muy simple. Supongamos que ni usted ni yo podemos distinguir un olmo de

un haya. Aún así, afirmamos que la referencia de «olmo» en mi discurso es idéntica a la referencia de «olmo» en el discurso de cualquier otra persona, v.gr. olmos, y que la extensión de «haya» es el conjunto de todas las hayas (es decir, el conjunto de cosas de las que la palabra «haya» se predica con verdad) tanto en su discurso como en el mío. ¿Es posible creer que la diferencia entre la referencia de «olmo» y la referencia de «haya» descansa en una diferencia en nuestros *conceptos?* Mi concepto de olmo es exactamente el mismo que mi concepto de haya (me ruboriza confesarlo). (Esto demuestra, de paso, que la determinación de la referencia es social y no individual; tanto usted como yo confiamos en los expertos que sí *pueden* distinguir los olmos de las hayas.) Si alguien intenta mantener heroicamente que la diferencia entre la referencia de «olmo» y la referencia de «haya» en *mi* discurso se explica por una diferencia en mi estado psicológico, dejémosle imaginar una Tierra Gemela en la que las palabras están trocadas. La Tierra Gemela es muy parecida a la Tierra; en realidad, aparte del hecho de que las palabras «olmo» y «haya» están intercambiadas, el lector puede suponer que la Tierra Gemela es exactamente como la Tierra. Supongamos que tengo un *doble* en la Tierra Gemela que es idéntico a mí, molécula por molécula, en el sentido en que dos corbatas pueden ser *idénticas.* Si usted es dualista, suponga que mi *doble* tiene los mismos pensamientos verbalizados que yo, experimenta los mismos datos sensoriales, posee las mismas disposiciones, etc. Es absurdo pensar que su estado psicológico es diferente del mío: pero aun así, su palabra «olmo» representa *hayas,* y mi palabra «olmo» representa olmos. (De forma similar, si el «agua» de la Tierra Gemela fuese un líquido diferente —*XYZ* y no H_2O, por ejemplo— entonces «agua» representaría líquidos distintos cuando se usa en la Tierra Gemela o cuando se usa en la Tierra.) Contrariamente a lo que afirma la doctrina que se ha mantenido entre nosotros desde el siglo diecisiete, *los significados no están en la cabeza.*

Hemos visto que poseer un concepto no consiste en poseer imágenes (de árboles, por ejemplo— o incluso de imágenes «visuales» o «acústicas» de oraciones, o de discursos enteros, dicho sea de paso), ya que alguien podría poseer cualquier sistema de imágenes que a usted se le antoje sin poseer por ello la *capacidad* de utilizar las sentencias de modos situacionalmente adecuados, considerando los factores lingüísticos —como afirmamos con anterioridad— y también a los no lingüísticos, como los determinantes de la «adecuación situacional». Un hombre podría tener todas las imágenes que usted quiera, y, sin embargo, no saber qué ha de hacer cuando se le dice «Señale el árbol», aun cuando haya presente un grupo de árboles. Hasta podría tener la imagen de lo que tiene que hacer, y, con todo, no saber lo que tiene que hacer. Pues la imagen, si no está acompañada por la capacidad de actuar de cierta manera, es sólo *imagen,* y la capaci-

dad de actuar de acuerdo «con un dibujo» es algo que se puede tener o no. Ese hombre podría representarse a sí mismo señalando un árbol, solamente con el ánimo de considerar algo lógicamente posible, después de que alguien produjese la secuencia de sonidos —que para él no tienen significado— «Por favor, señale un árbol». Y a pesar de todo no sabría que debe señalar un árbol, ni comprendería la exhortación «Señale un árbol».

He establecido que el criterio para poseer un «concepto efectivo» es la capacidad de usar ciertas oraciones, aunque este criterio puede ser liberalizado fácilmente. Podríamos permitir que los elementos del simbolismo no fuesen palabras del lenguaje natural, por ejemplo, y también podríamos dar entrada a fenómenos mentales como las imágenes y otras clases de eventos internos. Lo esencial es que éstos tengan la misma complejidad, la misma capacidad de ser combinados entre sí, etc., que las oraciones del lenguaje natural. Porque aunque una determinada representación —por ejemplo, un destello azul— podría servirle a un determinado matemático como expresión interior de toda la prueba del Teorema de los Números Primos, no tendríamos la tentación de afirmarlo (y sería falso hacerlo) si el matemático no pudiese desmembrar ese «destello azul» en pasos separados y conexiones lógicas. Pero, sin tener en cuenta cuáles son los tipos de fenómenos internos que permitimos como posibles *expresiones* del pensamiento, podrían ofrecerse argumentos exactamente similares al anterior que demostrarían que el entendimiento no está constituido por fenómenos, sino más bien por la capacidad del sujeto pensante para *emplear* y para producir fenómenos adecuados en las circunstancias adecuadas.

Acabo de ofrecer una versión muy abreviada del argumento de Wittgenstein de las *Investigaciones Filosóficas*. Si es correcto, entonces la pretensión de comprender el pensamiento a la luz de la denominada «investigación fenomenológica» se halla fundamentalmente desencaminada; pues lo que los fenomenólogos no ven es que lo que están describiendo es la *expresión interna* del pensamiento, pero la comprensión de esta expresión —la *comprensión* de los *pensamientos de uno mismo*— no es algo que acaezca, sino el resultado de una *capacidad*. Nuestro ejemplo de un hombre que simula pensar en japonés (y engaña a un telépata japonés) muestra de inmediato la falibilidad de una aproximación fenomenológica al problema de la *comprensión*. Pues aún habiendo alguna cualidad susceptible de introspección que estuviese presente cuando, y sólo cuando, uno *comprendiese realmente* (y esto, de hecho, parece ser falso con respecto a la introspección), esta cualidad estaría tan sólo *correlacionada* con la comprensión, y aún así sería posible que el hombre que engaña al telépata japonés se hallase en posesión de esa cualidad y, a pesar de todo, no entendiese ni una palabra de japonés.

Por otra parte, consideremos a un hombre (perfectamente posible) que no tiene ningún «monólogo interior». Habla perfectamente el castellano, y si le preguntamos cuáles son sus opiniones con respecto a un asunto dado, nos las ofrecerá detenidamente. Mas nunca piensa (con palabras, imágenes, etc.) cuando está hablando en voz alta; ni «pasa nada por su cabeza», salvo que (por supuesto) oye su propia voz hablando, tiene las impresiones sensoriales usuales de sus alrededores, más un «sentimiento general de comprensión». (Quizá tiene el hábito de hablar a solas.) Cuando mecanografía una carta, o cuando va de compras, etc., no experimenta un «flujo de pensamiento» pero sus acciones son inteligentes y propositivas, y si alguien le aborda y le pregunta «¿Qué está haciendo usted?», ofrecerá una respuesta perfectamente coherente.

Este hombre parece perfectamente imaginable. Nadie vacilaría en decir que es consciente, que no le gusta el rock and roll (si expresase con frecuencia una fuerte aversión por el rock and roll), etc., solamente porque no piensa conscientemente excepto cuando habla en voz alta.

De todo esto se sigue que (a) ningún conjunto de eventos mentales —imágenes o acontecimientos y cualidades más abstractas— *constituye* por sí mismo la comprensión; y (b) ningún conjunto de eventos mentales es *necesario* para la comprensión. En particular, *los conceptos no pueden ser idénticos a objetos mentales de ningún tipo.* Pues, aun suponiendo que entendemos por evento mental algo susceptible de introspección, cualquiera que sea éste, acabamos de ver que podría estar ausente en un hombre que entiende la palabra apropiada (y por tanto posee el concepto efectivo) y presente en un hombre que de ningún modo posee el concepto.

Volviendo ya a nuestra crítica de las teorías mágicas de la referencia (tópico que también interesaba a Wittgenstein), vemos que, por una parte, estos «objetos mentales» que *podemos* detectar introspectivamente —palabras, sentimientos, imágenes, etc.— no se refieren a algo intrínsecamente en mayor medida que lo hace el dibujo de la hormiga (y por las mismas razones), mientras que los intentos de postular objetos mentales especiales, «conceptos», los cuales *sí* tienen una conexión necesaria con sus referentes —y que sólo los fenomenólogos entrenados pueden detectar— incurren en un error *lógico,* porque los conceptos son (al menos en parte) *capacidades,* y no cosas que acontecen en la mente. La doctrina que defiende la existencia de representaciones mentales que se refieren necesariamente a las cosas externas no es sólo mala ciencia natural; es también mala fenomenología y confusión conceptual.

2. UN PROBLEMA ACERCA DE LA REFERENCIA

¿Por qué razón nos sorprende que la hipótesis de los cerebros en una cubeta resulte ser incoherente? La causa es nuestra tendencia a creer que lo que transcurre en el interior de nuestra cabeza debe determinar lo que queremos decir y aquello a lo que se refieren nuestras palabras. Pero no es difícil atisbar que esta creencia es errónea. Las palabras deícticas ordinarias, tales como *yo, éste, aquí, ahora,* constituyen triviales contraejemplos. Cuando pienso «Yo llego tarde al trabajo», puedo hallarme en el mismo estado mental en el que se halla Enrique cuando lo piensa (imagine el lector, si gusta, que Enrique y yo somos gemelos idénticos), y, con todo, la ocurrencia de la palabra «Yo» que se da en mi pensamiento se refiere a mí, y la que se da en el pensamiento de Enrique se refiere a Enrique. Cuando un martes pienso «Llego tarde al trabajo» puedo hallarme en el mismo estado mental [1] que cuando lo pienso un miércoles, pero el tiempo al que se refiere mi verbo «Llego», conjugado temporalmente, es diferente en cada uno de los casos. Los términos de géneros naturales ejemplifican más sutilmente el mismo punto.

Supongamos, ciñéndonos al caso mencionado en el capítulo anterior, que en la Tierra Gemela hay castellano-hablantes (por una especie de accidente milagroso, han evolucionado precisamente como nosotros y hablan un lenguaje idéntico al castellano que hace un par de siglos, dejando a un lado una diferencia que mencionaré de inmediato). Supondré que estas gentes no poseen todavía conocimientos de química daltoniana o postdaltoniana. De modo que, en particular, no disponen de nociones tales como «H_2O». Entonces, tal como se usa en la Tierra Gemela, la palabra «agua» no se refiere al agua, sino a este otro líquido (XYZ, digamos). Pese a todo, no hay diferencias

[1] Al menos puedo hallarme en el mismo estado mental en el sentido de que los parámetros implicados en el proceso psicológico que da por resultado mi pensar el pensamiento pueden tener los mismos valores. Seguramente mi estado mental global es distinto, ya que en martes pienso «Hoy es martes», y en miércoles no. Pero una teoría que afirmase que el significado de las palabras cambia cuando lo hace mi estado mental global, no permitiría que ninguna palabra conservase siempre el mismo significado, y ello equivaldría al abandono de la misma noción de significado de una palabra. Por otra parte, podríamos solapar aquí la historia de una Tierra Gemela, historia en la que mi doble y yo nos hallamos en el mismo estado mental global y aún así la referencia de «yo» y «ahora» sería con todo diferente (el calendario de la Tierra Gemela no está sincronizado con el nuestro).

relevantes entre el estado mental de los hablantes de la Tierra Gemela y el de los hablantes de la Tierra (en 1.750, digamos) que puedan dar cuenta de esta diferencia referencial. La referencia es diferente[2] porque la *sustancia* es diferente. Por sí mismo, aislado de la situación completa, el estado mental no fija la referencia.

Sin embargo, algunos filósofos han formulado objeciones a este ejemplo. Estos filósofos sugieren que, si tal planeta se descubriese alguna vez, uno diría que «Hay dos tipos de agua» y *no* que nuestra palabra «agua» no se refiere al líquido de la Tierra Gemela. De acuerdo con estos críticos, si encontrásemos alguna vez lagos y ríos cubiertos de un líquido que se parece superficialmente al agua, pero que no es H_2O, entonces habríamos falsado el enunciado de que toda agua es H_2O.

Es sencillo modificar el ejemplo con el fin de eludir este argumento. En primer lugar, el líquido de la Tierra Gemela no tiene por qué ser *tan* similar al agua. Supongamos que es realmente una mezcla de un 20 por 100 de alcohol de grano y otro 80 por 100 de agua, pero la composición química de la gente de la Tierra Gemela es tal que no se intoxican, ni siquiera notan la diferencia de sabor entre esa mezcla y el H_2O. Ese líquido sería diferente del agua en muchas cosas; con todo, un hablante *típico* podría no estar familiarizado con estas diferencias, y hallarse así exactamente en el mismo estado mental que un hablante *típico* de la Tierra en 1750. Para nosotros, desde luego, el «agua» de la Tierra Gemela tiene un gusto muy distinto al agua de la Tierra; pero para *ellos* no lo tiene. Por otra parte, cuando hierve reacciona de un modo distinto. Pero ¿acaso debe un castellano-hablante haber observado en qué momento exacto hierve el agua y qué ocurre exactamente, para asociar un contenido conceptual imparcial y estándar a la palabra «agua»?

Puede objetarse que bien podrían haber expertos en la Tierra Gemela que supieran cosas del «agua» (por ejemplo, que es una *mezcla* de dos elementos) que nosotros no sabemos sobre el agua (que no creímos sobre el agua en 1750, puesto que no son *verdaderas* del agua), y que, por consiguiente, el estado mental *colectivo* de los castellano-hablantes de la Tierra Gemela sería distinto del de los de la Tierra (en 1750). Uno podría conceder que el estado mental individual de una persona no fija la referencia de uno de sus términos, pero insistiendo en que lo hace el estado mental global de todos los miembros de la comunidad lingüística.

Esta objeción se topa con la dificultad de que la gente de la Tierra *o* de la Tierra Gemela podría no haber desarrollado tanta química en

[2] Para una discusión detallada de este punto, véase «The Meaning of "Meaning"» en mi libro *Mind, Language and Reality (Philosophical Papers,* vol. 2) Cambridge University Press, 1975.

1750. Si con anterioridad al desarrollo de la química, el término tenía en la Tierra el mismo significado y la misma referencia que hoy posee (en su uso ordinario) y si, con anterioridad al desarrollo de un conocimiento correspondiente, el término tenía en la Tierra Gemela el mismo significado y la misma referencia que hoy tiene, entonces podremos retroceder hasta ese tiempo anterior en el que los estados mentales colectivos de las dos comunidades *coincidían* en todos aquellos aspectos relevantes para fijar la extensión de «agua», argumentando que, pese a esto, la extensión era diferente *entonces* (como lo es ahora) y, de ahí, que el estado mental *colectivo* tampoco fija la extensión. ¿Hemos de decir que la referencia *cambió* cuando se perfeccionó la química? ¿Diremos que el término fue *usado* para referirse a ambos tipos de agua (a pesar de que para nosotros su *sabor* sea diferente) y que sólo se refiere a tipos *distintos* después de tal progreso?

Si decimos que cuando perfeccionaron (o perfeccionamos) la química (hasta el punto de poder destilar líquidos, de establecer que el agua más el alcohol componen una mezcla, etc...) cambió la referencia de sus términos (o de los nuestros) entonces tendremos que admitir que todos los descubrimientos científicos cambian la referencia de nuestros términos. Pero, según este punto de vista, nosotros no *descubrimos* que el agua (en el sentido precientífico) era H_2O; más bien lo *estipulamos*. Esta opinión me parece un craso error. Lo que nosotros queríamos decir con la palabra «agua» abarcaba desde un principio todo aquello que tuviese la misma naturaleza que la sustancia local *identificada* por ese término; y nosotros descubrimos que el agua, en ese sentido, era H_2O; lo que la gente de la Tierra Gemela quería decir con la palabra «agua» abarcaba desde un principio la sustancia de *su* entorno identificada por ese término, y sus expertos descubrieron que el «agua», *en ese sentido,* era una mezcla de dos líquidos.

Si estamos de acuerdo en que la palabra «agua» no cambia de *significado* (en uno u otro lenguaje) cuando los expertos llevan a cabo descubrimientos tales como «el agua es H_2O» o «el agua es una mezcla de dos elementos», o no cambia su significado y su referencia ordinarios (como resultado de tales descubriminetos pueden aparecer más usos técnicos, por supuesto), y que la palabra «agua», en su significado y su referencia ordinarios y terrestres, no incluye mezclas de alcohol y de agua, entonces debemos concluir que lo que da cuenta de la diferencia de *significado* del término «agua» en la Tierra y en la Tierra Gemela no es el conocimiento experto. Tampoco da cuenta de la *referencia:* todavía podríamos imaginar una segunda Tierra Gemela donde el agua fuese una mezcla *distinta* y el conocimiento experto fuera el mismo que en la Tierra Gemela (algo insuficiente). O, como ya indicamos, podemos imaginarnos que en la Tierra y en la Tierra Gemela todavía no existían expertos. La palabra «agua» se *referiría* a *una sustancia diferente* aun cuando el estado mental *colecti-*

vo de las dos comunidades fuese el mismo. Lo que ocurre en el interior de las cabezas de las personas no fija la referencia de sus términos. Con una frase debida a Mill, «la sustancia misma» completa la tarea de fijar la extensión del término.

Una vez vemos que el estado mental (en el sentido individual o colectivo) no fija la referencia, deja de sorprendernos el que los cerebros en una cubeta no logren referirse a objetos externos (aunque se hallen en nuestros mismos estados mentales), y que, por tanto, no puedan decir o pensar que son cerebros en una cubeta.

INTENCIONES, EXTENSIONES Y «MUNDOS NOCIONALES»

Puestos a examinar el problema de cómo se fija la referencia de nuestros términos, dado que no la fijan nuestros estados mentales, será conveniente disponer de algunos términos técnicos. En lógica, el conjunto de cosas de las que un término es verdadero se llama *extensión* del término. Así pues, la extensión del término «gato» es el conjunto de los gatos. Si un término tiene más de un sentido, entonces simulamos que el término lleva índices invisibles (de modo que tenemos realmente dos palabras y no una), por ejemplo «$conejo_1$» —extensión: conjunto de conejos; y «$conejo_2$» —extensión: conjunto de los cobardes. (Estrictamente hablando, la extensión de los términos en un lenguaje natural es siempre un poco difusa, pero fingiremos, por simplicidad, que los casos dudosos han sido legislados de algún modo.)

Una palabra como «yo», que se refiere a personas distintas en distintas ocasiones, no tendrá una extensión, sino una *función-extensión:* esto es, una función que determina una extensión en cada contexto de uso. En el caso de la palabra «yo», la función-extensión es bastante sencilla; es simplemente la función $f(x)$, cuyo valor para cualquier hablante x es el conjunto que consiste solamente en x. En semántica llamamos *índice* al argumento «x», que ranguea sobre todos aquellos parámetros relevantes usados en la descripción del contexto (en este caso, el *hablante*). Los índices son necesarios para el tiempo, para las cosas a las que nos referimos ostensivamente, y, en un tratamiento semántico completo, también para otras características del contexto (pero obviaremos los detalles).

El conjunto de cosas que componen la extensión de «gato» es distinto en las diferentes situaciones posibles o «mundos posibles». En un mundo posible M en el que no hay gatos, la extensión de «gato» es el conjunto vacío. Si mi gata Elsa hubiese tenido descendencia, entonces la extensión de «gato» tendría al menos un miembro que no tiene en el mundo actual. (Podemos expresar esto diciendo que en ca-

da mundo posible *M* en el cual Elsa tuvo descendencia, la extensión de «gato» incluye miembros que no incluye en el mundo actual.)

Podemos indicar la forma en que la extensión de un término varía con el mundo posible *M* del mismo modo que indicamos que la extensión de la palabra «yo» varía con el hablante: usando una *función*. Supondremos un conjunto de objetos abstractos, llamados «mundos posibles», que representan los diversos estados de hecho o posibles descripciones del mundo, y asociamos con el término «gato» una función *f(M)* cuyo valor en cada mundo posible *M* es el conjunto de objetos posibles que son gatos en el mundo *M*. Siguiendo a Carnap y Montague, llamaré a esta función *intensión*[3] de la palabra «gato». Similarmente, la intensión del predicado diádico «toca», es la función *f(M)*, cuyo valor en cada mundo posible *M* es el conjunto de pares ordenados de objetos posibles que en el mundo *M* se tocan entre sí; la intensión del predicado triádico «*x* está entre *y* y *z*» es la función cuyo valor en cualquier mundo posible es el conjunto de triplos ordenados *(x,y,z)* tales que *x* está entre *y* y *z*, y así sucesivamente. La intensión de una palabra como «yo», cuya extensión depende del contexto, en cada mundo posible, será una función más complicada que tendrá como argumentos los mundos posibles y los índices que representan el contexto.

La intensión especifica el modo en que la extensión depende del mundo posible. Representa pues lo que nos interesa, la extensión asociada a un término, de un modo muy completo, ya que nos dice qué extensión podría tener en cualquier mundo posible.

La razón por la que la «intensión» (en este sentido) no puede identificarse con el *significado*, es que cualesquiera dos términos lógicamente equivalentes coinciden extensionalmente en cada mundo posible, teniendo por tanto la misma intensión; pero una teoría que sea incapaz de distinguir entre términos que tienen el mismo significado y términos que son sólo lógica y matemáticamente equivalentes no es una teoría del *significado* adecuada. «Cubo» y «poliedro regular con seis caras cuadradas» son predicados lógicamente equivalentes. Así pues, la intensión de estos términos es la misma, a saber, la función cuyo valor en cualquier mundo posible es el conjunto de cubos en ese mundo. Pero hay cierta diferencia de significado que podría perderse si identificamos simplemente el significado con esta función.

Permítaseme poner énfasis en que, en esta teoría, los mundos posibles, los conjuntos y las funciones han de concebirse como entidades abstractas extramentales y no confundirse con representaciones o descripciones de estas entidades.

---

[3] R. MONTAGUE, *Formal Philosophy*, Yale University, 1974 *(Ensayos de Filosofía Formal*, Alianza, Madrid, 1977). Este uso de «intensión» no es el tradicional, discutido en «The Meaning of "Meaning"».

Frege pensó que el significado *(Sinn)* de una expresión era una entidad extramental o concepto, que podía ser «aprehendido» por la mente, no se sabe cómo. Esta teoría no nos sirve para el caso de las intensiones, entendidas en *nuestro* sentido. En primer lugar, como ya apunté, hay diferencias en significado que no pueden ser captadas por la intensión; por ende, comprender un término no puede consistir *sólo* en asociarlo con una intensión. Y lo que es más importante, si suponemos que no hay un «sexto sentido» que nos permita percibir *directamente* entidades extramentales (quizá «intuirlas»), entonces la «aprehensión» de una intensión, o de cualquier entidad extramental, debe de estar mediada de algún modo por representaciones. (Esto también parece claro introspectivamente, al menos para mí.) Pero el problema que estamos investigando es precisamente cómo las representaciones *pueden* permitir que nos refiramos a lo que se halla fuera de la mente. Dar por sentada la facultad de «aprehender» un x que es exterior a la mente sería cometer una petición de principio.

Al decir que alguien «cree que hay un vaso de agua en la mesa», lo que hago normalmente es atribuirle la capacidad de referirse al *agua*. Pero, como hemos visto, un requisito para poder referirnos al agua es estar directa o indirectamente vinculados con el agua real (H_2O); el enunciado «Juan cree que hay un vaso de agua delante de él», no versa sobre lo que ocurre en la cabeza de Juan, sino que es, en parte, un enunciado sobre el entorno de Juan y sobre su relación con ese entorno. Si resulta que Juan es un habitante de la Tierra Gemela, entonces lo que Juan cree cuando dice «Hay un vaso de agua en la mesa» es que en la mesa hay un vaso que contiene un líquido que *de hecho* consta de agua y alcohol de grano.

Husserl introdujo un recurso que es útil cuando deseamos hablar de lo que ocurre en nuestra cabeza sin tener que admitir ninguna presuposición con respecto a la existencia o la naturaleza de las cosas reales a las que se refieren los pensamientos: el recurso de la *puesta entre paréntesis*[4]. Si «ponemos entre paréntesis» la creencia que adscribimos a Juan cuando decimos «Juan cree que hay un vaso de agua en la mesa», entonces simplemente le adscribimos el *estado mental* de una persona real o posible que cree que hay un vaso de agua en la mesa (en el pleno sentido ordinario y «sin poner entre paréntesis»). De modo que si Juan no puede notar la diferencia de sabor entre el agua y el agua-con-alcohol-de-grano de la Tierra Gemela, puede hallarse en el mismo estado mental que un hablante real o posible de la Tierra cuando dice «Hay un vaso de agua en la mesa», a pesar del hecho de que Juan se refiera con la palabra «agua» a algo que consti-

4 HUSSERL, *Ideas; General Introduction to Pure Phenomenology*, Allen and Unwin, 1969.

tuiría un cocktail algo cargado. Diremos que tiene la *creencia «puesta entre paréntesis»* de que [hay un vaso de agua en la mesa]. En efecto, el recurso de la puesta entre paréntesis resta las implicaciones de la locución ordinaria de creencia (todas las implicaciones que se refieren al mundo externo, o a lo que es exterior a la mente del sujeto pensante).

Daniel Dennett ha utilizado recientemente la locución «mundo nocional» de un modo afín a como Husserl utilizó la puesta entre paréntesis [5]. La totalidad de las creencias puestas entre paréntesis del sujeto pensante constituye la descripción de su mundo nocional, en el sentido de Dennett. Así pues, la gente de la Tierra Gemela tiene aproximadamente el mismo mundo *nocional* (e incluso la misma agua *nocional*) que tenemos nosotros; lo único que ocurre es que viven un planeta *real* diferente y (usan «agua» para referirse a una sustancia real distinta); y los cerebros en una cubeta del capítulo precedente podrían haber tenido el mismo mundo nocional que nosotros tenemos, hasta el último detalle; lo que pasa es que ninguno de sus términos tenía referencia en el mundo exterior. La teoría tradicional del significado daba por sentado que el mundo nocional del sujeto pensante determinaba las intensiones de sus términos (y éstas, junto con el hecho de que un mundo posible particular M es el actual, determinan las extensiones de los términos y los valores veritativos de las oraciones). Hemos visto que la teoría tradicional del significado es incorrecta; por eso la literatura contiene hoy muchos conceptos diferentes (por ejemplo «intensión» y «mundo nocional»), y no un solo concepto unitario de «significado». El «significado» *se ha hecho pedazos.* Pero nos queda la tarea de recoger éstos. Si la intensión y la extensión no están directamente determinadas por el mundo nocional, ¿cómo están determinadas entonces?

LA CONCEPCION ADMITIDA CON RESPECTO A LA INTERPRETACION

La opinión más común con respecto al modo en que fijamos las interpretaciones de nuestro lenguaje (si no individual, al menos colectivamente), está asociada con las nociones de *constreñimiento operacional* y *constreñimiento teórico.* Los constreñimientos operacionales se concibieron originariamente de un modo bastante ingenuo; simplemente *estipulamos* (convencionalmente, como si fuera así) que cierta oración (digamos «Fluye electricidad a través del alambre») es verda-

---

[5] DENNETT, D., «Beyond Belief», en *Thought and Object,* Andrew Woodfield (ed.), Oxford University Press, 1984.

dera si y solo si se observa cierto resultado en una prueba. (La oscilación de la aguja del voltímetro, o, en lenguaje fenomenalista, *mi impresión visual de ver oscilar la aguja del voltímetro.*) Esta especie de crudo operacionalismo ya no tiene defensores, puesto que se ha apreciado que (1) los vínculos entre teoría y experiencia son probabilísticos y no pueden ser correctamente formalizados como perfectas correlaciones (aun cuando fluya electricidad a través del alambre, hay siempre sucesos de baja probabilidad o condiciones de fondo que podrían impedir la oscilación de la aguja del voltímetro); y (2) estos vínculos ni siquiera son simples correlaciones *semánticas,* sino que dependen de una teoría empírica que está sujeta a revisión. Según el operacionalismo ingenuo, cada vez que se descubre un nuevo test para comprobar si una sustancia es realmente «oro», el significado y la referencia de oro sufre un cambio. (De hecho, no diríamos que se *descubre* un nuevo criterio para reconocer el oro.) Desde una perspectiva operacionalista, las teorías son contrastadas con la experiencia *oración por oración* (los significados operacionales estipulados para las oraciones individuales nos dicen cómo emprender la contrastación de la teoría); según una descripción más reciente, las teorías se someten al tribunal de la experiencia como cuerpos integrados, como expuso Quine.

Sin embargo, es posible relajar la noción del constreñimiento operacional con el fin de superar estas objeciones. Así, uno podría restringir la clase de interpretaciones (asignaciones de intensiones a los predicados de su lenguaje) admisibles, de acuerdo con constreñimientos de la forma: «una interpretación es admisible si la *mayoría de las veces la oración S* es verdadera cuando se satisface la condición experimental *E*» (respectivamente, si la mayoría de las veces la *oración S* es *falsa* cuando se satisface *E*). Tales constreñimientos conforman la idea de que hay relaciones probabilísticas entre la experiencia y la verdad o la falsedad de las oraciones del lenguaje. Y, en segundo lugar, se puede adoptar la opinión de que estos constreñimientos son revisables conforme se desarrolla la teoría. En vez de concebirlos como estipulaciones de significado, a la manera del operacionalismo crudo, pueden concebirse como tentativas de restricción de la clase de interpretaciones admisibles; y adoptar con Pierce (que escribió 50 años antes que Bridgeman proclamara el «operacionalismo») la opinión de que el conjunto ideal de constreñimientos operacionales es en sí mismo algo a lo que nos vamos aproximando sucesivamente en el curso de la indagación empírica, y no algo que estipulemos. En resumen, puede adoptarse la opinión de que lo que selecciona la interpretación de nuestros términos son los constreñimientos operacionales que los investigadores racionales *impondrían,* si observasen y experimentasen razonando tan correctamente como fuera posible: los constreñimientos que adoptarían en el estado de «equilibrio reflexivo»; los constre-

ñimientos que de hecho aceptamos en un momento dado tienen el *status* de estimación racional o aproximación.

Tal opinión es compatible con la insistencia quineana en que los vínculos teoría-experiencia se hallan tan sujetos a revisión como cualquier otro aspecto de nuestro cuerpo integrado de conocimiento. Y ello no significa considerar cada revisión como un «cambio de significado»: tales revisiones pueden ser, y a menudo son, simples esfuerzos para especificar aquello de lo que ya hemos estado hablando; la teoría más tosca de los constreñimientos operacionales sólo captaba este punto inadecuadamente.

Además de restringir la clase de interpretaciones admisibles por medio de constreñimientos operacionales (o mediante aproximaciones sucesivas al conjunto ideal peirciano de constreñimientos operacionales), también podemos disponer de constreñimientos que se remitan a propiedades formales de la teoría. Por ejemplo, «Una interpretación admisible es aquella en la que resulta ser cierto que diferentes efectos tienen diferentes causas». Kant sostuvo que tal «constreñimiento teórico» formaba parte de la propia racionalidad: *imponemos* en el mundo el principio del determinismo, más bien que descubrirlo. Esta formulación del principio es demasiado enfática: el *precio* de preservar el determinismo podría ser una complicación excesiva de nuestro sistema cognitivo, considerado como un todo. Pero este tipo de constreñimientos puede relajarse, tal y como lo han sido los constreñimientos operacionales. (Por ejemplo, podemos exigir que se preserve el determinismo sólo cuando el «costo», ponderado en términos de las complicaciones que ocasiona en otras partes de nuestro sistema cognitivo, no es excesivo. Así formulado, este constreñimiento sí parece ser uno de los que aceptamos.)

Aunque a menudo se formulan los constreñimientos teóricos como constreñimientos para la aceptación de teorías más bien que como constreñimientos para su interpretación, pueden reinterpretarse fácilmente para que desempeñen este último papel. De este modo, aunque un autor formule el constreñimiento de «conservadurismo» o «preservacionismo» a modo de límite para la aceptación teórica («no aceptemos una teoría que requiera abandonar un gran número de creencias previamente aceptadas, si disponemos de otra teoría —por otra parte, igualmente «simple»— que las preserva y que concuerda además con la observación») podemos reformularlo para que desempeñe un papel análogo en la interpretación teórica, del siguiente modo: «Una interpretación admisible es aquella que proporciona oraciones verdaderas que se aceptan por un largo período de tiempo, excepto donde ello requiera una complicación indebida de la teoría —consistente en el conjunto de oraciones verdaderas bajo la interpretación— o una revisión excesiva de los constreñimientos operacionales». Comúnmente se acepta, de nuevo, que no es posible ningún tipo de lógica inductiva

a no ser que impongamos algún orden *a priori* (llamado «orden de simplicidad» u «orden de plausibilidad») sobre las hipótesis que pueden aceptarse a partir de unos datos determinados (si bien el propio orden puede ser diferente en distintos contextos experimentales y observacionales); el constreñimiento: «El conjunto de oraciones verdadero bajo una interpretación no debe tener un grado inferior de simplicidad que cualquier otro conjunto que posea las mismas consecuencias observacionales y experimentales», correspondería en lógica inductiva al constreñimiento que conmina a aceptar la hipótesis más simple (o la más «plausible») entre aquellas que son compatibles con las observaciones que se llevan a cabo.

En la literatura de la filosofía de la ciencia se han propuesto constreñimientos teóricos de muchos otros tipos; hay algunos, como el de «simplicidad», que remiten a propiedades del conjunto de oraciones aceptadas, y otros que aluden en cambio a la *historia* de la investigación por la que tal conjunto se aceptó finalmente. Pero no necesitamos entretenernos con los detalles. Los motivos que hacen atractiva la idea de que las interpretaciones de nuestro lenguaje (en el sentido de asignaciones de intensiones a sus términos) se fijan por medio de constreñimientos operacionales y teóricos, son obvios: la mente puede discernir si se está teniendo o no cierto tipo de experiencia (a pesar de los problemas filosóficos anejos a la «experiencia»). De forma que si una teoría implica o contiene una oración asociada con una experiencia *E* por medio de algún tipo de constreñimiento operacional, probabilístico, o de cualquier otra índole, entonces el sujeto pensante puede saber si la teoría es operativa, o si, por el contrario, hay alguna dificultad de ajuste —al menos en ese caso— observando si tiene o no la experiencia *E*. Y dado que los constreñimientos utilizados para contrastar la teoría fijan también las extensiones de los términos, lo que el sujeto pensante estime sobre la «operatividad» de la teoría constituye también una estimación sobre la *verdad* de ésta. Y ya que el hablante, al conocer esos constreñimientos, conoce también las intensiones de los términos, se sigue que la aprehensión de una semántica correcta le informaría de cómo ha de ser el mundo para que la teoría —cualquier teoría del mundo propuesta— sea verdadera.

Además, si idealizamos mediante la suposición de que cada sujeto pensante posee lo que los economistas llaman «información perfecta» acerca de los demás sujetos pensantes, cada uno conocerá la estructura formal de la teoría *T* aceptada, la historia del programa de investigación al cual pertenece, las creencias previas que preserva o no, etc. De este modo, cada sujeto pensante se hallaría en una posición óptima para *saber* si se satisfacen o no los constreñimientos teóricos. (Si no deseamos la idealización consistente en suponer tal información perfecta, podríamos decir, a pesar de todo, que es el

cuerpo colectivo de sujetos pensantes el que está en posición de saberlo.)

En resumen, si la concepción admitida fuese correcta, tendríamos una elegante descripción de cómo se fijan las intensiones y las extensiones (en principio, pues es difícil atar todos los cabos, dado el estado actual de nuestro conocimiento metodológico). Pero, por desgracia, la concepción admitida no resuelve el problema.

POR QUE FRACASA LA CONCEPCION ADMITIDA

La dificultad de la concepción admitida es que intenta fijar las intensiones y las extensiones de los términos individuales fijando las condiciones de verdad de las oraciones completas. Como ya vimos, la idea es que los constreñimientos operacionales y teóricos (aquellos que los investigadores racionales aceptarían en algún tipo de límite ideal de investigación) determinan qué oraciones son *verdaderas* en el lenguaje. Sin embargo, aunque esto sea correcto, tales constreñimientos no pueden determinar la *referencia* de nuestros términos. No hay nada en la noción de constreñimiento operacional o teórico que permita determinarla directamente. Y hacerlo *indirectamente,* preservando los constreñimientos que identifican el conjunto de oraciones verdaderas, y esperando entonces que mediante la determinación de los valores de verdad de oraciones completas podremos fijar de algún modo la referencia de los *términos* que ocurren en éstas, no resolverá el problema.

Tal fracaso ha sido mostrado por Quine[6]. Voy a prolongar radicalmente sus resultados previos con respecto a la «indeterminación». Argumentaré que incluso disponiendo de constreñimientos —cualquiera que sea su naturaleza— que determinen el valor de verdad de cada oración de un lenguaje en cada *mundo posible,* la referencia de los términos individuales aún permanece indeterminada. De hecho, es posible interpretar violentamente todo un lenguaje de diferentes maneras, cada una de ellas compatible con el requisito de que el valor de verdad de cada oración en cada mundo posible sea el especificado. Resumiendo, no sólo es que fracase la concepción admitida: *ningún criterio que únicamente fije los valores de verdad de oraciones completas puede fijar la referencia, incluso si especifica los valores de verdad de las oraciones en cada mundo posible.*

La prueba pormenorizada es técnica y creo que lo apropiado es ofrecerla en un Apéndice . Lo que ofreceré aquí es sólo una ilustra-

[6] «Ontological Relativity», en *Ontological Relativity and Other Essays,* Columbia University Press, 1969 *(La Relatividad Ontológica y Otros Ensayos,* Tecnos, Madrid, 1974).

ción del método de la prueba, y no la prueba detallada. Consideremos la oración:

(1) Un gato está en una estera. (De aquí en adelante, «está en» es *atemporal,* es decir, significa «está, estaba, o estará en».)

Bajo la interpretación estándar, esta oración es verdadera en aquellos mundos posibles en los cuales hay al menos un gato y al menos una estera en algún tiempo, pasado, presente o futuro. Además «gato» se refiere a gatos y «estera» a esteras. Mostraré que puede reinterpretarse la oración (1) de manera que, en el mundo actual, «gato» se refiera a *cerezas* y «estera» a *árboles,* sin que ello afecte al valor veritativo de (1) en cada mundo posible. («Está en» conservará su interpretación original.)

La idea es que la oración (1) reciba una nueva intepretación en la cual llegue a significar:

(2) Un gato* está en una estera*.

La definición de la propiedad de ser un gato* (y, respectivamente, una estera*), viene dada por casos, siendo los tres casos:

(a) Algún gato está en alguna estera y alguna cereza en algún árbol.
(b) Algún gato está en alguna estera y ninguna cereza en ningún árbol.
(c) Ninguno de los precedentes.

He aquí la definición de las dos propiedades:

DEFINICION DE «GATO*»

x es un gato* si y solo si se cumple el caso (a) y *x* es una cereza; o se cumple el caso (b) y *x* es un gato; o se cumple el caso (c) y *x* es una cereza.

DEFINICION DE «ESTERA*»

x es una estera si y solo si se cumple el caso (a) y x* es un árbol; o se cumple el caso (b) y *x* es una estera; o se cumple el caso (c) y *x* es un quark.

Ahora bien, en los mundos posibles que caen bajo el caso (a), «Un gato está en una estera» es verdadera, y también lo es «Un gato* está en una estera*», (porque una cereza está en algún árbol, y en los mundos de este tipo todas las cerezas son gatos* y todos los árboles son esteras*). Ya que en el mundo actual alguna cereza está en algún ár-

bol, el mundo actual es un mundo de este tipo, y en él, «gato*» se refiere a cerezas y «estera*» se refiere a árboles.

En los mundos posibles que caen bajo el caso (b), «Un gato está en una estera» es verdadera; y también lo es «Un gato* está en una estera*» (porque en los mundos posibles que caen bajo el caso (b), «gato» y «gato*» son términos coextensivos, como lo son «estera» y «estera*»). (Démonos cuenta de que aunque los gatos son gatos* en algunos mundos —los que caen bajo el caso (b)— no son gatos* en el mundo *actual.*)

En los mundos posibles que caen bajo el caso (c), «Un gato está en una estera» es falsa y «Un gato* está en una estera*» también lo es (porque una cereza no puede estar en un *quark*).

En suma, vemos que en *cada mundo posible* un gato está en una estera si y solo si un gato* está en una estera*. De este modo, la reinterpretación de la palabra «gato», asignándole la intensión que precisamente le asignamos a «gato*», y la reinterpretación simultánea de la palabra «estera», asignándole la intensión que precisamente le asignamos a «estera*», únicamente tendría el efecto de hacer que «Un gato está en una estera» signifique lo mismo que lo que «Un gato* está en una estera*» significa por definición; y ello sería perfectamente compatible con el modo en el que se asignan valores veritativos a la oración «Un gato está en una estera» en cada mundo posible.

En el Apéndice muestro que puede llevarse a cabo una reinterpretación de este tipo, pero más complicada, para todas las oraciones de un lenguaje completo. Se sigue que siempre hay un número infinito de diferentes interpretaciones de un lenguaje que asignan a las oraciones los valores de verdad «correctos» en todos los mundos posibles, *sin que importe el modo en que se especifiquen tales valores veritativos.* Quine abogó por una conclusión semejante en *Palabra y Objeto;* en el ejemplo de Quine (aplicado al castellano) «Hay un conejo allá» se reinterpretaba para que significase «Hay una porción de conejo allá» (donde una «porción de conejo» es una sección transversal, espacial y tridimensional del todo espacio-temporal y cuatridimensional que es el conejo) o, alternativamente, «¡Oh, conejidad otra vez!». (Esta última reinterpretación también reinterpreta la forma sintáctica de la oración, o al menos su gramática lógica.) Quine establece el punto que acabo de establecer, esto es, que *las condiciones de verdad* para *oraciones completas* subdeterminan la referencia. Ya que «porción de conejo», «conejidad» y «partes no-separadas de conejo» están estrechamente relacionadas con conejos, uno podría salir de *Palabra y Objeto* con la impresión de que todas las reinterpretaciones que no alteran el valor veritativo de una oración están al menos estrechamente relacionadas con la interpretación estándar (del mismo modo que las «partes de conejo» y la «conejidad» están relacionadas con conejos). El argumento detallado en el Apéndice e ilustra-

do en este capítulo demuestra que las condiciones de verdad para «Un gato está en una estera» ni siquiera excluyen la posibilidad de que «gato» se refiera a *cerezas*.

«INTRINSECO» Y «EXTRINSECO»

Quizá lo primero que se nos ocurre cuando hacemos frente a interpretaciones no-estándar, tal y como la que interpreta «gato» como gato* y «estera» como estera*, es desecharlas, ya que nos presentan una insignificante paradoja. Pero las paradojas *genuinas* nunca son insignificantes; revelan siempre alguna incorrección en nuestra forma de pensar. Quizá la segunda reacción sea denunciar que gato* y estera* son propiedades «excéntricas»; nuestros términos corresponden a propiedades sensatas (tales como *ser un gato* o *ser una estera*) y no a propiedades «irrisorias» como éstas, por supuesto. Podríamos aclarar por qué gato* (o mejor *gatidad** o *gat*idad*) es una propiedad extraña, señalando que podemos «construir una máquina» que «examine cosas» y «diga» si son o no gatos (el ser humano es una «máquina» de tal guisa), pero no podemos construir una máquina que nos diga (en cualquier mundo cuyas leyes y condiciones generales sean parecidas a las del nuestro) si algo es o no un gato*. Si la máquina (o una persona) escruta algo y observa que ni es un gato ni una cereza, entonces puede decir que *no es un gato**; pero si ese algo es un gato o una cereza, entonces el artilugio o la persona necesita ser informado de los valores veritativos de «Un gato está en una estera» y «Una cereza está en un árbol» para decidir si está examinando u observando un gato*; pero estos valores de verdad exceden lo que puede aprender mediante el mero examen del objeto que se le presenta.

Por desgracia, uno puede reinterpretar «observa» (por ejemplo, como observa*) de tal modo que las dos oraciones (3) Juan (o quien sea) observa un gato y (4) Juan observa* un gato*, tengan el mismo valor de verdad en cada mundo posible (por el método ofrecido en el Apéndice). De modo que cuandoquiera que una persona observa un gato, está observando* un gato*; la experiencia que tenemos característicamente cuando observamos un gato es la experiencia que tenemos característicamente cuando observamos* un gato* y así sucesivamente. Podemos reinterpretar de forma similar «examina» y «dice», de modo que cuando una máquina «examina» un gato, está examinando* un gato*, y cuando «dice» que algo es un gato, está diciendo* que es un gato*.

Para valernos de un ejemplo (sugerido por Nozick), supongamos que la mitad de nosotros (quizá las mujeres) utiliza «gato» para significar «gato*», «estera» para significar «estera*», «mirar» para

«mirar*», «decir» para «decir*». Supongamos que la otra mitad (los hombres) usa «gato» para denotar gatos, «estera» para denotar esteras, etc. ¿Cómo podríamos llegar a saberlo? [7]. (Si preguntamos a un hombre a qué se refiere con «gato», responderá «a gatos, por supuesto» y así lo hará una mujer, *sea cual sea* la referencia de «gato».)

Lo importante aquí es que el hecho de que tengamos la posibilidad de *construir una máquina que examine cosas y diga si son gatos,* sólo diferencia a los gatos de los gatos* si podemos estar seguros que «examine» y «diga» se refieren a examinar y decir, y no es más fácil indicar cómo se fija la referencia de estas palabras que indicar cómo se fija la de «gato». Podría argumentarse que cuando miro algo y pienso que es un gato, mis «representaciones mentales», mis imágenes visuales o táctiles, mi pensamiento verbalizado «gato», etc., se *refieren* a la gatidad y a otras propiedades físicas o biológicas (ser cierta forma, cierto color, pertenecer a cierta especie) y no a sus duplicados. Pese a que esta réplica podría ser acertada, reincide en suponer que la referencia se fija más bien de un modo que de otro. Esto es precisamente lo que deseamos explicar, y no la explicación solicitada.

«Pero» podría alguien protestar, «las definiciones de «gato*» y «estera*» dadas anteriormente se refieren a cosas que no son los objetos en cuestión (cerezas sobre árboles y gatos sobre esteras) y significan pues propiedades extrínsecas de los objetos que las poseen. En el mundo actual, cada cereza es un gato*; pero aunque sus propiedades intrínsecas fueran exactamente las mismas, no sería un gato* si no hubiese ninguna cereza en ningún árbol. Por contraste, que algo sea o no un gato depende únicamente de sus propiedades intrínsecas. La distinción aquí mencionada, la distinción entre propiedades *intrínsecas* y *extrínsecas,* ¿nos permite caracterizar y rechazar las interpretaciones «extrañas»?

El problema de esta sugerencia es cierta simetría en la relación de «gato» y «estera» con gato* y estera*. Supongamos que definimos «cereza*» y «árbol*», de forma que en los mundos posibles que caen bajo el caso (a) las cerezas* son gatos y los árboles* son esteras; en los mundos que caen bajo el caso (b) las cerezas* son cerezas y los árboles* son árboles; y en los mundos posibles que caen bajo el caso (c) las cerezas* son gatos y los árboles* son fotones. Entonces, podemos definir «gato» y «estera» por medio de *-términos, tal y como sigue: Casos:

---

[7] Una mujer podría responder que la suposición de que se esté refiriendo a gatos* cuando dice «gato» es *incoherente* (porque *dentro* de su lenguaje cualquier cosa a la que se refiere con «gato» es un gato). Esta respuesta es sólo un pequeño consuelo; no excluye la posibilidad de que lo que *ella* llama *gato* es lo que un hombre llama *gato** y viceversa; y éste es el chiste de Nozick.

(a*) Algún gato* está en alguna estera*, y alguna cereza* está en algún árbol*.
(b*) Algún gato* está en alguna estera* y no hay ninguna cereza* en ningún árbol*.
(c*) Ninguno de los precedentes.

Aunque parezca bastante extraño, estos casos son precisamente nuestros antiguos casos (a), (b) y (c) bajo una nueva descripción. Ahora definimos:

DEFINICION DE «GATO»

x es un gato si se cumple el caso (a*) y *x* es una cereza*; o se cumple el caso (b*) y *x* es un gato*; o se cumple el caso (c*) y *x* es una cereza*. (Admitamos que, en los tres casos, los gatos acaban siendo *gatos*.)

DEFINICION DE «ESTERA»

x es un estera si y solo si se cumple el caso (a*) y *x* es un árbol*; o se cumple el caso (b*) y *x* es una estera*; o se cumple el caso (c*) y *x* es un quark* (suponiendo que quark* se defina de modo que en los casos de tipo (c*) los quark* son *esteras*, las esteras acaban siendo *esteras* en los tres casos).

El resultado es que, vistas desde la perspectiva de un lenguaje que tome «gato*» y «estera*», etc., como propiedades primitivas, «gato» y «estera» se refieren a propiedades «extrínsecas», cuya definición menciona objetos que no son *x;* mientras que relativamente a un lenguaje «normal», un lenguaje que tome «gato» y «estera» para referirse a la gatidad y a la esteridad (usted ya sabe a qué propiedades me refiero, querido lector), «gato*» y «estera*» se refieren a propiedades «extrínsecas». Mejor dicho, ser «intrínseco» o «extrínseco» es algo relativo a las propiedades que uno elige como *básicas;* ninguna propiedad es en sí misma intrínseca o extrínseca.

«SUPERVIVENCIA» Y EVOLUCION

La sugerencia que hoy en día está de moda es que el mismo proceso evolutivo ha producido, de un modo u otro, una correspondencia entre nuestras palabras, las representaciones mentales y las cosas externas; se dice que no habríamos *sobrevivido* si no se hubiese dado tal correspondencia, y que ésta constituye la relación de referencia, por lo menos en un nivel rudimentario.

Pero ¿qué tienen que ver «correspondencia» y «referencia» con la *supervivencia*? Y, de paso, ¿qué tiene que ver la *verdad* con la supervivencia?

Aquí las opiniones difieren. Algunos filósofos creen que no sobreviviríamos si (una cantidad suficiente de) nuestras creencias no fuesen *verdaderas*. Otros afirman que ni siquiera nuestras creencias científicas mejor establecidas son verdaderas, o al menos que no tenemos razón alguna para pensar que lo son. Thomas Kuhn ha sugerido que nuestras creencias sólo se «refieren» a objetos *dentro* de esas creencias (de un modo parecido a como «Hamlet» únicamente se refiere a una persona en una obra); el éxito de la ciencia se explica por ensayo-y-error y no por una correspondencia entre sus objetos y cosas reales, apunta Kuhn. En su nuevo libro, Bas van Fraasen sostiene que una teoría exitosa no necesita ser verdadera, sino sólo «observacionalmente adecuada», esto es, que prediga correctamente la observación. También explica el éxito (o la adecuación observacional) de la ciencia como producto del ensayo-y-error.

Si estos filósofos están en lo cierto, la idea de emplear la evolución para justificar la creencia en una relación *objetiva* de referencia es gratuita. Según la perspectiva instrumentalista, la evolución únicamente establece una correspondencia entre algunos términos (los términos observacionales) y «posibilidades» permanentes de sensación. Tal correspondencia no puede ser la *referencia,* a menos que deseemos abandonar la idea de que las cosas externas (las observables) son algo más que meros constructos a partir de sensaciones.

No obstante, creo que son los otros filósofos quienes están en lo cierto (aquéllos que dicen que no sobreviviríamos si una cantidad suficiente de nuestras creencias no fuesen *verdaderas*).

Y lo creo por una razón: el ensayo-y-error no explica por qué nuestras teorías son «observacionalmente adecuadas»; *tal explicación* sólo puede darse en relación con las características de la *interacción* hombre-medio, siendo éstas las que explican el *éxito del ensayo-y-error.* (¡El ensayo-y-error no tiene éxito en *todas* las empresas, después de todo!) Postular que la interacción ocasiona en nuestras mentes teorías *falsas* cuyas consecuencias son precisamente predicciones exitosas es postular una serie de *coincidencias* totalmente inexplicable. Pero ¿cómo puede explicarse nuestra supervivencia mediante el hecho de que nuestras creencias son (aproximadamente) verdaderas?

Algunas de nuestras creencias están íntimamente ligadas con la *acción.* Si creo la oración «Si pulso este botón conseguiré algo de gran valor para mí» (suponiendo que entiendo esta oración de un modo normal, o al menos le asocio la creencia «puesta entre paréntesis» o «nocional» normal), entonces alargaré la mano y pulsaré el botón. Llamo *creencias directivas* a las creencias de la forma: «Si hago x, conseguiré...», donde el espacio en blanco representa una *meta* del

agente. De modo que la verdad de (una cantidad suficiente de) nuestras creencias *directivas* es necesaria para la supervivencia.

Ahora bien, nuestras creencias *directivas* se derivan a su vez de otras muchas creencias: creencias sobre las características y sobre las facultades causales de las cosas externas, y creencias sobre nuestras propias características y facultades. Si la mayor parte de estas creencias fuesen falsas, ¿no sería una mera coincidencia que, pese a ello, condujesen a predicciones de la experiencia y a creencias directivas verdaderas? Así pues, ya que (una cantidad suficiente de) nuestras creencias directivas son verdaderas, y puesto que la *mejor explicación* de este hecho es que muchas de nuestras restantes creencias (las que constituyen nuestra «teoría del mundo cotidiano») son verdaderas, al menos aproximadamente, tenemos buenas razones para creer que nuestra teoría del mundo cotidiano es verdadera, al menos aproximadamente, y que si no fuera así no hubiésemos sobrevivido.

Imaginemos ahora que algunos de nosotros nos referimos realmente a las cosas que la interpretación *no-estándar J* (descrita en el Apéndice) asigna a nuestros términos. Esta interpretación coincide con la *estándar* en los términos referentes a nuestro mundo nocional, nuestras creencias, voliciones, etc. De esta manera, la oración «Me parece que he pulsado el botón», si la entendemos «puesta entre paréntesis» (significando que tengo cierta experiencia subjetiva de haber pulsado voluntariamente el botón), no sólo tiene las mismas condiciones de verdad, sino también la misma *interpretación* bajo *J* y bajo la interpretación «normal» *I;* y lo mismo ocurre con la oración «Me parece que he obtenido la satisfacción que esperaba».

Ahora bien, si una cantidad suficiente de nuestras creencias directivas son verdaderas bajo la interpretación no-estándar *J*, entonces, desde luego, obtendremos éxito y *sobreviviremos* (ya que si no estuviéramos vivos no estaríamos alcanzando esas metas) y tendremos descendencia (ya que si *ellos* no estuvieran vivos, no estarían alcanzando esas metas). En resumen, la *J*-verdad de (una cantidad suficiente de) creencias directivas es tan óptima para el «éxito evolutivo» como la *I*-verdad, ya que las condiciones de verdad de *cada oración* (no sólo de las creencias directivas) son las mismas bajo *I* y bajo *J*. Mis creencias directivas no *sólo* están asociadas con la misma *experiencia* subjetiva bajo la interpretación *I* y bajo la interpretación *J;* tienen además las mismas *condiciones de verdad.* Desde el punto de vista (o no-punto de vista) de la «evolución», todo lo que se necesita es que una cantidad suficiente de mis creencias sean verdaderas bajo *cualquier* interpretación que conecte tales creencias con *acciones* relevantes. La evolución puede provocar en mí cierta tendencia a tener *creencias verdaderas* (de ciertos tipos); pero esto únicamente quiere decir que la evolución influye lingüística o conceptualmente sobre la supervivencia *mediante* su tendencia a generar en nosotros sistemas de represen-

tación cuyas oraciones, o análogos de oración, poseen ciertas *condiciones de verdad* (y ciertas condiciones de *acción* o «reglas de salida del lenguaje»). *Pero ya mostramos que las condiciones de verdad para oraciones completas no determinan la referencia de sus partes* (tampoco nos ayudará «añadir reglas de salida del lenguaje» puesto que éstas se preservan bajo *J*). Se sigue que es sencillamente un error pensar que la evolución determina una *única* correspondencia (o incluso un rango razonablemente restringido de correspondencias) entre las expresiones referenciales y los objetos externos.

INTENCIONES: PURAS E IMPURAS

Hemos visto que la naturaleza no selecciona ninguna correspondencia entre nuestros términos y las cosas externas. La naturaleza nos conmina a tratar las palabras y los signos mentales de tal modo que una cantidad suficiente de nuestras creencias directivas resulten verdaderas, a fin de que una cantidad suficiente de nuestras acciones contribuyan a nuestra «aptitud genética completa»; pero esto deja la referencia en gran parte indeterminada. W. V. Quine ha insistido en que esto es, de hecho, lo que la referencia es: indeterminada. El autor de *La Relatividad Ontológica* piensa que es una quimera considerar que los términos de nuestro lenguaje tengan contrapartidas bien-definidas. Con sus palabras:

> Consideremos de nuevo nuestra notación canónica, con un léxico de predicados interpretados y algún rango de valores fijado para las variables de cuantificación. Las oraciones verdaderas de este lenguaje continúan siendo verdaderas bajo incontables reinterpretaciones de los predicados e incontables revisiones del rango de valores de las variables. En realidad, podemos hacer que sirva cualquier rango del mismo tamaño por medio de una adecuada reinterpretación. Si el rango de valores es infinito, podemos hacer que sirva cualquier rango infinito; éste es el Teorema de Skolem-Löwenheim. Las oraciones verdaderas continúan siendo verdaderas bajo todos estos cambios.
> Quizá entonces nuestra principal preocupación tenga que ver, más que con la referencia de los términos, con la verdad y las condiciones veritativas de las oraciones.

La alternativa aquí sugerida, y que examinaré en el próximo capítulo, es abandonar la idea que hasta aquí ha servido como premisa de toda la discusión: que nuestras palabras están en una especie de relación uno-a-uno con cosas (independientes del discurso) y con conjuntos de cosas. Sin embargo, puede parecer que hay una solución mucho más simple: ¿por qué no decir que son precisamente nuestras *intenciones*, implícitas o explícitas, las que fijan la referencia de nuestros términos?

Nada más comenzar la discusión en el capítulo previo, rechacé esta opción sobre la base de que no constituía una respuesta informati-

va, puesto que el tener intenciones (del tipo relevante) presupone ya la capacidad de referir. En este punto puede que convenga ampliar ese breve comentario.

El problema es que las nociones de «intención» y «estado mental» poseen cierta ambigüedad. Llamemos *puro* a un estado mental si su presencia o ausencia depende sólo de lo que acaece dentro del hablante. El que yo tenga o no un dolor sólo depende de lo que acaece «dentro de mí», pero que sepa o no que la nieve es blanca depende no sólo de que algo acaezca o no «dentro» de mí (creer o confiar en que la nieve es blanca), sino que también depende de que la nieve sea o no blanca, y, de este modo, de algo que acaece «fuera» de mi cuerpo y de mi mente. Así pues, el *dolor* es un estado mental puro, pero el *conocimiento* es un estado mental *impuro*. Hay un componente mental (puro) en el conocimiento, pero también hay un componente que de ningún modo es mental: aquél que corresponde a la condición de que lo que un hombre cree no constituye conocimiento a menos que la creencia sea *verdadera*. No me hallo en el «estado» de conocer que la nieve es blanca si no estoy en un estado mental puro conveniente. Pero estar en un estado mental puro conveniente *nunca basta para saber que la nieve es blanca;* el mundo tiene que cooperar también.

¿Y la creencia? Hemos definido la creencia *puesta entre paréntesis* («mundo nocional») de forma que tener una creencia puesta entre paréntesis de que [hay agua en la mesa], o tener un mundo nocional que incluya el haber agua sobre la mesa, es un estado mental puro. Pero, de acuerdo con lo que decíamos anteriormente, *el creer que hay agua en la mesa* (sin poner nada «entre paréntesis») presupone que la palabra «agua» *se refiere realmente al agua,* y esto depende de la naturaleza real de ciertos «paradigmas», de las relaciones causales directas que se mantengan con estos paradigmas, etc. Cuando tengo la creencia de que hay agua en la mesa, mi doble de la Tierra Gemela tiene la misma creencia puesta entre paréntesis, pero no la misma creencia, ya que su palabra «agua» se refiere al agua-con-alcohol-de grano y no al agua. En resumen, creer que hay agua en la mesa es un estado mental *impuro* (los cerebros en una cubeta no podrían estar en tal estado, aunque sí en el correspondiente estado «puesto entre paréntesis»).

A la intención le ocurre lo mismo que a la creencia. Los estados mentales e intencionales puros —por ejemplo, el acto intencional de que el término «agua» se refiera al agua *en el mundo nocional de uno*— no fijan la referencia en el mundo real. Los estados mentales intencionales impuros —el acto intencional de que el término «agua» se refiera al agua real— *presuponen* la capacidad de referirse al agua (real).

Algunos filósofos han sugerido que la creencia puede definirse en términos del estado que llamé «creencia puesta entre paréntesis» *y* de la referencia, de este modo:

> Juan cree que la nieve es blanca = *Juan cree que [la nieve es blanca].*

(Esto es, la nieve es blanca en el mundo nocional de Juan.)

> *Y en el pensamiento de Juan, las palabras «nieve» y «blanca» (o cualesquiera palabras que use para expresar su creencia) se refieren a la nieve y a la propiedad de ser blanca, respectivamente.*

Sin admitir esta descripción como un análisis correcto y completo de lo que es creer que la nieve es blanca, podemos aceptar que pone de relieve un punto que sin duda es correcto: *creer presupone la capacidad de referir.* Y exactamente del mismo modo, los actos intencionales ya presuponen la capacidad de referir. Las intenciones no son acontecimientos mentales que *causan* la referencia de las palabras: las intenciones (en el sentido ordinario e impuro) contienen la referencia como un *componente* integral. Explicar la referencia en términos de intención (impura) sería circular. Y el problema de cómo los estados mentales *puros* intencionales, de creencia, etc., pueden (en el marco causal apropiado) constituir o causar la referencia es precisamente lo que hemos hallado demasiado enigmático.

EL ORIGEN DEL ENIGMA

A primera vista parece sumamente obvio que nuestras palabras y representaciones mentales *refieren.* Cuando pienso o digo «El gato acaba de salir», usualmente pienso algo acerca de nuestro gato Mitty; la palabra gato, en esa oración, pensada o dicha, se *refiere* a un conjunto de entidades del que es miembro Mitty. A pesar de todo, acabamos de ver que la naturaleza de esta relación de «acerca de qué» o referencia es enigmática.

La distinción entre el mundo real y el mundo nocional (y la correspondiente distinción entre creencias y creencias puestas entre paréntesis, o entre intenciones e intenciones puestas entre paréntesis) explica parte del enigma. La razón por la que resulta sorprendente y problemático descubrir que hay «interpretaciones admisibles» no intencionales de nuestro lenguaje (entiendo por interpretación admisible una interpretación que satisface los constreñimientos operacionales y teóricos apropiados) es, en parte, que en el «mundo nocional» del hablante no surge tal «indeterminación». En mi mundo nocional los gatos y los gatos* son completamente distintos (de hecho, en mi mundo nocional los gatos* son *cerezas*). «Hay un gato en una estera» y «Hay un gato* en una estera*» pueden ser lógicamente equivalen-

tes, pero contienen términos con referentes nocionales totalmente distintos; por eso parece realmente extraño que pueda haber confusión entre los referentes del mundo real de una creencia y los de la otra.

Pero si el número de gatos resulta ser igual que el número de cerezas, entonces se sigue de los teoremas de la Teoría de Modelos (como Quine subraya en la cita anterior) que existe una reinterpretación de todo el lenguaje que no altera los valores veritativos de todas las oraciones aunque se permuten las extensiones de «gato» y «cereza». Gracias a las técnicas ya mencionadas, tales interpretaciones pueden concebirse de manera que preserven todos los constreñimientos operacionales y teóricos (y mediante las técnicas ejemplificadas en el caso «gato/gato*», pueden ampliarse hasta que nos proporcionen «intensiones», o funciones que determinen una extensión *en cada mundo posible,* y no sólo extensiones en el mundo actual). Ello no contradice los enunciados que establecimos sobre nuestro mundo nocional, o sistema subjetivo de creencias, por la siguiente razón: el hecho de que en nuestro sistema de creencias o «mundo nocional» ningún gato sea una cereza significa que en cada interpretación de tal sistema (en cada asignación de referentes del mundo externo a los términos, a las imágenes y a las demás representaciones que empleamos al pensar), los referentes de «gato» y los referentes de «cereza» deben ser conjuntos disjuntos. Pero la disyuntividad de tales conjuntos es comparable con el hecho (subrayable) de que lo que es el conjunto de los «gatos» en *una* interpretación admisible puede ser el de las «cerezas» en una interpretación *diferente* (pero igualmente admisible). Del hecho de que los gatos nocionales sean distintos de las cerezas nocionales sólo se sigue que los gatos reales son completamente distintos de las cerezas reales si el número de interpretaciones es exactamente uno. Si hay más de una interpretación admisible de todo el lenguaje (como las habrá si las interpretaciones admisibles son seleccionadas exclusivamente por los constreñimientos operacionales y teóricos), entonces dos términos que se refieren a conjuntos disjuntos en *cada* interpretación admisible, pueden tener los mismos referentes potenciales cuando se tiene en cuenta la totalidad de interpretaciones admisibles. Del hecho de que los gatos nocionales sean diferentes de las cerezas nocionales no se sigue que haya conjuntos disjuntos de gatos-en-sí-mismos y cerezas-en-sí-mismas; y si esto resulta tan penoso es porque los constreñimientos operacionales *más* los teóricos son el medio natural para permitir que el contexto empírico determine la interpretación (o las interpretaciones) admisibles del sistema representacional de un sujeto. Tales constreñimientos pueden determinar, hasta cierto punto, las oraciones que son verdaderas y las que son falsas en nuestro lenguaje; y ésta es la escasa actividad que queda entre las condiciones de verdad y la referencia. Como ya señalamos, Quine desearía acabar con esta relación moribunda y reconocer simplemente que la referencia

queda indeterminada. Hartry Field [8], un joven filósofo, ha sugerido una alternativa distinta. En opinión de Field, la referencia es una «relación fisicalista», es decir, una relación causal compleja entre palabras o representaciones mentales y objetos o conjuntos de objetos. Según Field, corresponde a la ciencia empírica descubrir lo que sea tal relación fisicalista. No obstante, esta sugerencia también tropieza con problemas. Supongamos que, como Field sostiene, hay una posible *definición* fisicalista o naturalista de la referencia.

Supongamos que:

(1) *x se refiere a y* si y solo *si x mantiene R con y* es verdadera, donde *R* es una relación definible con el vocabulario de la ciencia natural, sin usar nociones semánticas (es decir, sin usar «refiere» o cualesquiera otras palabras que harían la definición inmediatamente circular). Si (1) es verdadera y empíricamente verificable, entonces (1) es una oración verdadera hasta en la teoría según la cual la referencia se fija en tanto que (y *sólo* en tanto que) es determinada por los constreñimientos operacionales *más* los teóricos. Así pues (1) es una oración que formaría parte de nuestra teoría del mundo en el «límite ideal» o «equilibrio reflexivo».

Sin embargo, si la referencia sólo está determinada por los constreñimientos operacionales y teóricos, entonces la referencia de «*x mantiene R con y*» está *ella misma* indeterminada, y por ende, saber que (1) es verdadera será de poca ayuda. Cada modelo admisible de nuestro lenguaje-objeto tendrá su correspondiente modelo en nuestro metalenguaje, en el cual (1) se cumpla. La interpretación de «*x mantiene R con y*» fijará la interpretación de «x se refiere a y». Pero ésta será sólo una relación *en cada modelo admisible;* de ninguna manera servirá para acotar el número de interpretaciones admisibles.

Field no pretende afirmar esto, desde luego. Field está afirmando que (a) hay una relación única entre palabras o cosas o conjuntos de cosas; y (b) ésta es la relación que debemos usar como relación de referencia al asignar un valor veritativo a (1). Pero, como acabamos de ver, no expresamos necesariamente esto sólo con *decir* (1); y cómo podríamos *aprender a expresar* que lo que Field quiere decir es un misterio.

Dejando a un lado este último enigma, consideremos la opinión de que (1), si la entendemos como Field quiere que lo hagamos (como si describiese la relación única y determinada entre las palabras y sus referentes), es verdadera. Si, entendida de este modo, (1) es verdade-

[8] FIELD, H., «Tarski's Theory of Truth», *The Journal of Philosophy,* vol. 69. El punto de vista de Field se discute en mi libro *Meaning and the Moral Sciences,* Routledge and Kegan Paul, 1978.

ra, ¿qué la *hace* verdadera? Dado que hay muchas «correspondencias» entre palabras y cosas, e incluso muchas correspondencias que satisfacen nuestros constreñimientos, ¿qué *selecciona* a una particular correspondencia *R*? No es la corrección empírica de (1); porque esto es asunto de nuestros constreñimientos operacionales y teóricos. No son, como hemos visto, nuestras intenciones (más bien *R* toma parte en la determinación de lo que nuestras intenciones significan). Parece que el hecho de que *R sea* la referencia debe ser un hecho *metafísicamente inexplicable,* un tipo de verdad metafísica primitiva, irracional.

Este tipo de verdad metafísica primitiva, irracional, si la hubiese, no debería confundirse con la clase de verdad «metafísicamente necesaria», dada a conocer recientemente por Saul Kripke[9].

La indicación de Kripke, que está íntimamente relacionada con las anteriormente realizadas sobre la referencia de los términos de géneros naturales (por ejemplo, términos para especies animales, vegetales y minerales), consistía en que, *dado* que como cuestión de hecho

(2) El agua es H_2O.

(es decir, dado que (2) es verdadera en el mundo real) y dado que (señala Kripke) los hablantes *efectúan el acto intencional* de que el término «agua» se refiera sólo a aquellas cosas que tienen la misma conducta legal y la misma composición última que los diversos estándares del agua real (esto es, los hablantes tienen tales intenciones incluso cuando hablan sobre casos hipotéticos o «mundos posibles»), se sigue que (2) también debe ser verdadera en todo mundo posible; porque describir un hipotético líquido que no es H_2O, pero que tiene algunas características similares a las del agua, es sólo describir un hipotético líquido que *se asemeja* al agua, y no describir un mundo posible en el que el agua no es H_2O. Es «metafísicamente necesario» (verdad en todos los mundos posibles) que el agua es H_2O; pero esta «necesidad metafísica» se explica por la química mundana y por los hechos mundanos sobre las intenciones referenciales de los hablantes.

Incluso si hay una relación fisicalista determinada *R* (sea o no definible en el lenguaje de la ciencia con un número finito de palabras) que *sea* precisamente la referencia (independientemente de cómo la describamos), *este* mismo hecho no puede ser consecuencia de nuestras intenciones referenciales; más bien, tal y como hemos apuntado repetidas veces, toma parte en la determinación de lo que nuestras mis-

[9] Véase su libro *Naming and Necessity,* Harvard University Press, 1980.

mas intenciones referenciales significan. La opinión de Kripke según la cual «El agua es H_2O» es verdadera en todos los mundos posibles, podría ser acertada incluso si la referencia en el mundo real se fijase sólo por constreñimientos operacionales y teóricos; su punto de vista presupone la noción de referencia, no nos dice si la referencia está determinada o qué es la referencia.

En mi opinión, creer que alguna correspondencia es precisa e intrínsecamente la referencia (no como resultado de nuestros constreñimientos operacionales y teóricos, o de nuestras intenciones, sino como un hecho metafísico *último*) viene a ser lo mismo que mantener una teoría mágica de la referencia. Desde tal perspectiva, la propia referencia se convierte en lo que Locke llamó «forma substancial» (una entidad que pertenece *intrínsecamente* a cierto nombre). Aun cuando estemos deseando contemplar tales hechos metafísicos inexplicables, los problemas epistemológicos que acompañan a tal perspectiva metafísica parecen insuperables. Porque, suponiendo un mundo de entidades independientes de la mente y del discurso (y esto es lo que presupone el punto de vista que estamos discutiendo) hay, como hemos visto, muchas «correspondencias» diferentes que representan relaciones de referencia posibles o candidatas (infinitamente muchas, de hecho, si hay una cantidad infinita de cosas en el universo). Ni siquiera la exigencia de que (1) sea verdadera, bajo la noción de verdad que corresponda a la relación «real» de referencia metafísicamente seleccionada, logra excluir ninguna de estas candidatas, si (1) es *empíricamente* aceptable (aceptable según nuestros constreñimientos operacionales y teóricos). Mas entonces hay una cantidad infinita de posibles «verdades metafísicas irracionales» *diferentes,* de la forma «*R* es la relación de referencia real (metafísicamente singularizada)». Si el que sostiene tal punto de vista concede que podría no ser del todo correcto, y que la referencia puede estar metafísicamente seleccionada sin estar totalmente *determinada* (la relación metafísicamente seleccionada *R* puede permitir una pluralidad de interpretaciones admisibles), entonces hasta puede ser concebible que el punto de vista de los constreñimientos-operacionales-y-teóricos sea metafísicamente correcto, después de todo. Pues ¿por qué no podría ser un hecho metafísico inexplicable el que la referencia sea la relación: *x* se refiere a *y al menos en un modelo admisible M?* Démonos cuenta de que *todas* estas infinitas teorías metafísicas son compatibles con las *mismas* oraciones verdaderas, con la misma «teoría del mundo» y con la misma metodología óptima para descubrir lo que es verdadero.

3. DOS PERSPECTIVAS FILOSOFICAS

Los problemas que hemos estado discutiendo dan origen de por sí a dos puntos de vista filosóficos (o a dos temperamentos filosóficos, tal y como los llamé en la introducción). Lo que me interesa son estos puntos de vista y sus consecuencias con respecto a cada problema filosófico: la cuestión de los «cerebros en una cubeta» carecería de interés, salvo como una especie de paradoja lógica, si no fuera por la nitidez con que subraya las diferencias entre estas perspectivas filosóficas.

Una de ellas es la del realismo metafísico. Según esta perspectiva, el mundo consta de alguna totalidad fija de objetos independientes de la mente. Hay exactamente una descripción verdadera y completa de «cómo es el mundo». La verdad supone una especie de relación de correspondencia entre palabras o signos mentales y cosas o conjuntos de cosas externas. A esta perspectiva la llamaré *externalista,* ya que su punto de vista predilecto es el del Ojo de Dios.

La perspectiva que voy a defender carece de nombre que no sea ambiguo. Es un logro tardío en la historia de la filosofía, y todavía hoy se preocupa de que no se la confunda con otros puntos de vista de índole completamente distinta. La denominaré perspectiva *internalista,* ya que lo característico de tal concepción es sostener que sólo tiene sentido formular la pregunta *¿de qué objetos consta el mundo?* desde *dentro* de una teoría o descripción. Muchos filósofos internalistas, aunque no todos, sostienen además que hay más de una teoría o descripción del mundo verdadera. Desde la perspectiva internalista, la «verdad» es una especie de aceptabilidad racional (idealizada) —una especie de coherencia ideal de nuestras creencias entre sí y con nuestras experiencias, *considerándolas como experiencias representadas en nuestro sistema de creencias*— y no una correspondencia con «estados de cosas» independientes de la mente o del discurso. No existe un punto de vista como el del Ojo Divino que podamos conocer o imaginar con provecho. Sólo existen diversos puntos de vista de personas reales, que reflejan aquellos propósitos e intereses a los que se subordinan sus descripciones y teorías. («Teoría de la verdad-coherencia», «no realismo», «verificacionismo», «pluralismo», «pragmatismo»; pese a que se han aplicado todos estos términos a la perspectiva internalista, cada uno de ellos tiene connotaciones inaceptables debido a sus restantes aplicaciones históricas).

Los filósofos internalistas rechazan la hipótesis de los «cerebros en una cubeta». La hipótesis de los «cerebros en un mundo-cubeta»

es para nosotros únicamente un *relato,* una mera construcción lingüística: de ningún modo un mundo posible. La idea de que este relato podría ser verdadero en algún universo, en alguna Realidad Paralela, supone desde el principio el punto de vista del Ojo de Dios, como fácilmente puede verse. En efecto, *¿desde qué punto de vista se cuenta este relato?* Evidentemente, *no desde* el punto de vista de alguna criatura sintiente en el mundo. Tampoco desde el punto de vista de algún observador de otro mundo que interactúe con éste, pues un «mundo» incluye por definición todo aquello que interactúa, de una u otra forma, con las cosas que contiene. Si *usted,* por ejemplo, fuera el observador que *no* es un cerebro en una cubeta, espiando a los cerebros en una cubeta, entonces el mundo no sería un mundo en el que *todos* los seres sintientes fueran cerebros en una cubeta. Así que la suposición de que podría haber un mundo en el que todos los seres sintientes fueran cerebros en una cubeta presupone desde el principio la visión de la verdad del Ojo Divino —o, con más precisión, la visión de la verdad del No-Ojo—, la verdad como algo totalmente independiente de los observadores.

Por otra parte, el filósofo externalista no puede rechazar tan fácilmente la hipótesis de que todos somos cerebros en una cubeta. Pues la verdad de una teoría no consiste en su ajuste con el mundo conforme éste se presenta al observador u observadores (la verdad no es «relacional» en este sentido), sino en su correspondencia con el mundo tal como es en sí mismo. Y el problema que le planteo al filósofo externalista, es que si él es un cerebro en una cubeta, no puede disponer lógicamente de la misma relación de correspondencia de la que (en su opinión) dependen la verdad y la referencia. Por tanto, si *somos* cerebros en una cubeta, no podemos *pensar* que lo somos, excepto en el sentido puesto entre paréntesis [Somos cerebros en una cubeta]. Y este pensamiento puesto entre paréntesis no tiene condiciones referenciales que lo hagan *verdadero.* De modo que, después de todo, no es posible que seamos cerebros en una cubeta.

Supongamos que asumimos una «teoría mágica de la referencia». Podríamos suponer, por ejemplo, que algunos rayos ocultos —llamémosles «rayos noéticos» [1]— conectan las palabras y los símbolos mentales con sus referentes. No hay problema entonces. El cerebro en una cubeta puede pensar las palabras «Soy cerebro en una cubeta», y cuando lo hace la palabra «cubeta» corresponde (con la ayuda de los rayos noéticos) a cubetas externas reales, y la palabra «en» (con idéntica ayuda) a la relación espacial de contención. Pero tal opinión es obviamente insostenible. Ningún filósofo de nuestros días se adheriría a ella. El caso de los cerebros en una cubeta es un problema para

[1] Zemach me sugirió los «rayos noéticos».

el moderno realista debido a su deseo de poseer una teoría de la verdad-correspondencia *sin* tener que creer en «rayos noéticos» (o sin tener que creer en objetos que se Auto-Identifican[2]—Objetos que corresponden intrínsecamente a una palabra o signo mental más bien que a otro).

Como hemos visto, el problema es el siguiente: ahí fuera hay esos objetos. Aquí la mente-cerebro, llevando a cabo su pensamiento-computación. ¿Cómo entran los símbolos del sujeto pensante (o los de su mente-cerebro) en una correspondencia única con los objetos y conjuntos de objetos de ahí fuera?

La réplica que está hoy de moda entre los externalistas es que, pese a que en realidad ningún signo corresponde *necesariamente* a ningún conjunto de cosas más bien que a otro, las conexiones *contextuales* entre los signos y las cosas externas (en particular, las conexiones causales) harán explicable la naturaleza de la referencia. Pero así no se resuelve el problema. Por ejemplo, es probable que la causa predominante de mis creencias sobre los electrones sean diversos *manuales*. Pero mis proferencias de la palabra «electrón», pese a tener en este sentido una firme conexión con los manuales, no se *refieren* a éstos. Los objetos que son la causa predominante de mis creencias contenientes de cierto signo pueden no ser los referentes de ese signo.

El externalista replicará ahora que la palabra «electrón» no está conectada con los manuales mediante una cadena causal del *tipo apropiado*. (Pero ¿cómo podemos tener intenciones que determinen qué cadenas causales son «del tipo apropiado» salvo que *ya* seamos capaces de *referirnos* a las cosas?)

Para un internalista como yo, la situación es completamente distinta. Desde una perspectiva internalista, los signos tampoco corresponden intrínsecamente a objetos con independencia de quién y cómo los emplee. Pero un signo empleado de un modo determinado por una determinada comunidad de usuarios puede corresponder a determinados objetos *dentro del esquema conceptual de esos usuarios*. Los «objetos» no existen independientemente de los esquemas conceptuales. Desmenuzamos el mundo en objetos cuando introducimos uno u otro esquema descriptivo, y puesto que tanto los objetos como los símbolos son internos al esquema descriptivo, es posible indicar cómo se emparejan.

En realidad es trivial decir cuál es la referencia de alguna palabra *dentro* del lenguaje al que pertenece mediante el uso de la misma palabra. ¿A qué se refiere «conejo»? ¡Cómo! ¡A conejos, por supuesto! ¿A qué se refiere extraterrestre? A extraterrestres (si los hay).

Por supuesto, el externalista está de acuerdo en que la extensión

---

[2] El término «Objeto qué se Auto-Identifica» proviene de *Substance and Sameness* de DAVID WIGGINS, Blackwell, 1980.

de «conejo» es el conjunto de los conejos y la de «extraterrestre» el conjunto de extraterrestres. Pero no considera que tales enunciados nos indiquen qué *es* la referencia. Para él, averiguar qué *es* la referencia, es decir, cuál es la *naturaleza* de la correspondencia entre las palabras y las cosas, *es* un problema apremiante (y en el capítulo anterior vimos *cuán* apremiante). Para mí, hay poco que decir sobre lo que es la referencia dentro de un esquema conceptual, aparte de estas tautologías. La idea de que es necesaria una conexión causal es refutada por el hecho de que «extraterrestre» se refiere sin duda a extraterrestres, hayamos interactuado causalmente con extraterrestres o no.

No obstante, el filósofo externalista replicaría que podemos referirnos a extraterrestres pese a no haber interactuado con ninguno de ellos (que sepamos) gracias a que hemos interactuado con *terrestres* y a que hemos experimentado instancias de la relación «no es del mismo planeta que» e instancias de la propiedad «ser inteligente». Y podemos *definir* extraterrestre como ser inteligente que no es del mismo planeta que los terrestres. Además «no es del mismo plantea que» puede analizarse en términos de «no es del mismo lugar que» y «planeta» (que pueden seguir analizándose). El externalista renuncia de este modo al requisito de que tengamos alguna conexión real (por ejemplo, una conexión causal) con *todas las cosas* a las que podemos referirnos, exigiendo sólo que los términos *básicos* se refieran a géneros de cosas (y relaciones) con los que tenemos alguna conexión real. El externalista afirma que utilizando los términos básicos en combinaciones complejas podemos construir expresiones descriptivas que se refieran a géneros de cosas con las que no tenemos ninguna conexión real, y que pudieran incluso no existir (por ejemplo, extraterrestres).

Lo cierto es que podría haberse dado cuenta de que la extensión de palabras tan simples como «conejo» o «caballo» incluye muchas cosas con las que *no* hemos interactuado causalmente (por ejemplo, conejos y caballos *futuros,* o conejos y caballos que nunca interactuaron con seres humanos). Cuando usamos la palabra «caballo» no sólo nos referimos a los caballos con los que tenemos conexión real, sino también a todas las demás cosas del *mismo tipo.*

No obstante, debemos observar en este punto que «del mismo tipo que» es una expresión que no tiene sentido si no es desde un sistema categorial que señale qué propiedades cuentan y qué propiedades no cuentan como semejanzas. Después de todo, cualquier cosa es del mismo tipo que cualquier otra de *varias* maneras. Todo este alambicado relato acerca de cómo nos referimos a algunas cosas en virtud del hecho de que estén conectadas con nosotros por cadenas causales del tipo apropiado, y aún a otras en virtud del hecho de que son «del mismo tipo que» las cosas conectadas con nosotros por cadenas causales del tipo apropiado, e incluso a otras por «descripción», no es

tan falso como ocioso. Lo que hace que los caballos con los que no he interactuado sean del «mismo tipo» que aquéllos con los que sí lo he hecho, es que tanto los primeros como los últimos son *caballos*. La formulación del problema por parte del realista metafísico lo disfraza una vez más, como si hubiera que empezar con todos esos objetos en sí mismos, adquiriendo entonces un tipo peculiar de lazo con algunos pocos (los caballos con los que tengo conexión real, *vía* cadena causal del tipo apropiado), topándome entonces con el problema de conseguir que mi palabra («caballo») cubra no sólo aquellos objetos que he «enlazado», sino también aquellos otros que no puedo enlazar, bien porque están demasiado lejos en el espacio-tiempo, bien por cualquier otra causa. Y la «solución» a este pseudo-problema, tal como yo lo considero —la «solución» del realista metafísico— es decir que la palabra cubre *automáticamente* no sólo los objetos que he enlazado, sino también los objetos que son *del mismo tipo* —del mismo tipo en *sí mismos*. Pero entonces lo que se afirma, después de todo, es que el mundo consta de Objetos que se Auto-Identifican, pues afirmar que es el *mundo,* y no los sujetos pensantes, el que clasifica las cosas en géneros, significa precisamente esto.

Yo diría que el mundo *sí* consiste en Objetos que se Auto-Identifican en un sentido —pero en un sentido no asequible al externalista. Si, como mantengo, los propios objetos son tanto construidos como descubiertos, son tanto producto de nuestra invención conceptual como del factor «objetivo» de la experiencia, el factor independiente de nuestra voluntad, entonces los objetos pertenecen intrínsecamente a ciertas etiquetas; porque esas etiquetas son las herramientas que usamos para construir una versión del mundo en la que tales objetos ocupan un lugar preeminente. Pero *este* tipo de Objetos que se Auto-Identifican no es independiente de la mente; y lo que el externalista quiere es concebir el mundo como si consistiese de objetos que son independientes de la mente y que *al mismo tiempo* se Auto-Identifican. Y esto es lo que no se puede hacer.

INTERNALISMO Y RELATIVISMO

El internalismo no es un fácil relativismo que afirme que «todo vale». Negar que tenga sentido preguntar si nuestros conceptos «se emparejan» con algo completamente incontaminado por la conceptualización es una cosa. Pero inferir a partir de esto que cualquier esquema conceptual es tan bueno como cualquier otro sería otra muy distinta. Si alguien creyese algo semejante, y fuera lo bastante imprudente como para escoger un esquema conceptual que le dijese que pue-

de volar y que obre en consecuencia saltando por la ventana, observaría de inmediato (si tuviera la suerte de sobrevivir) las desventajas del último punto de vista. El internalismo no niega que haya *inputs* experienciales en el conocimiento; el conocimiento no es un relato que no tenga otra constricción que la coherencia *interna;* lo que niega es que existan *inputs* que no *estén configurados en alguna medida por nuestros conceptos,* por el vocabulario que utilizamos para dar cuenta de ellos y para describirlos, o *inputs que admitan una sola descripción, independiente de toda opción conceptual.* Hasta la descripción de nuestras propias sensaciones, tan estimada como punto de partida del conocimiento por generaciones de epistemólogos, está profundamente afectada (como lo están las mismas sensaciones, dicho sea de paso) por multitud de opciones conceptuales. Los propios *inputs* sobre los que se basa nuestro conocimiento están conceptualmente contaminados. Pero es mejor tener *inputs* contaminados que no tener *inputs* de ninguna clase. Y si todo lo que tenemos son *inputs* contaminados, aun así no tenemos poco.

Lo que hace que un enunciado, o un sistema completo de enunciados —una teoría o esquema conceptual— sea racionalmente aceptable es, en buena parte, su coherencia y ajuste; la coherencia de las creencias «teóricas» —o menos experienciales— entre sí y con las creencias más experienciales; y también la coherencia de las creencias experienciales con las teóricas. Según la teoría que voy a desarrollar, nuestras concepciones de coherencia y aceptabilidad están profundamente entretejidas en nuestra psicología. Dependen de nuestra biología y de nuestra cultura y no están, en absoluto, «libres de valores». Pero *son* nuestras concepciones, y lo son de algo real. Definen un tipo de objetividad, *objetividad para nosotros,* si bien ésta no es la objetividad metafísica del punto de vista del Ojo de Dios. Objetividad y racionalidad humana es lo que tenemos; y tener esto es mejor que no tener nada.

Rechazar la idea de una perspectiva «externa» coherente, una teoría que simplemente es verdadera en sí misma, dejando a un lado todo posible observador, no es *identificar* la verdad con la aceptabilidad racional. La verdad no puede *ser* tan sólo aceptabilidad racional por una razón fundamental; se supone que la verdad es una propiedad perenne de un enunciado, mientras que la justificación puede perderse. Con toda probabilidad, el enunciado «La tierra es plana» era racionalmente aceptable hace 3.000 años, pero no lo es hoy. No obstante, sería erróneo decir que «La tierra es plana» era *verdadero* hace 3.000 años, ya que ello significaría que la tierra ha cambiado de forma. En realidad, la aceptabilidad racional es relativa tanto a un tiempo como a una persona. Es además una cuestión de grado; a veces la verdad es presentada como una cuestión de grado (por ejemplo decimos a veces que *«La tierra es una esfera» es aproximadamente ver-*

dadero; pero aquí se trata del grado de *precisión* del enunciado, y no de su grado de aceptabilidad o justificación).

En mi opinión, esto no muestra que la perspectiva externalista sea al fin y al cabo correcta, sino que la verdad es una *idealización* de la aceptabilidad racional. Hablamos como si hubiera tales cosas como condiciones epistemológicas ideales, y llamamos «verdadero» a un enunciado que estaría justificado bajo tales condiciones. Las condiciones epistemológicamente ideales son como las superficies sin rozamiento: en realidad no podemos obtener condiciones epistemológicamente ideales, ni siquiera tener la certeza de que nos hemos aproximado suficientemente a ellas. Pero tampoco podemos conseguir superficies sin rozamiento, y aún así decimos que las superficies sin rozamiento tienen «valor efectivo» gracias a que podemos acercarnos a ellos con un grado de aproximación bastante alto.

Quizá parezca que explicar la verdad en términos de justificación bajo condiciones ideales es explicar una noción clara en términos de otra vaga. Pero la «verdad» *no es* tan clara cuando nos alejamos de ejemplos tan banales como «La nieve es blanca». Y, en cualquier caso, no estoy intentando ofrecer una *definición* formal, sino una elucidación informal, de la noción de verdad.

Dejando a un lado el símil de las superficies sin rozamiento, las dos ideas clave de la teoría de la verdad-idealización son las siguientes: (1) la verdad es independiente de la justificación aquí y ahora, no independiente de *toda* justificación. Afirmar que un enunciado es verdadero es afirmar que podría ser justificado. (2) Es de esperar que la verdad sea estable o «convergente»; si tanto un enunciado como su negación pueden ser «justificados», no tiene sentido pensar que tal enunciado *posee* un valor de verdad, por mucho que las condiciones fueran tan ideales como uno soñase alcanzar.

LA TEORIA DE LA «SIMILITUD»

La teoría que afirma que la verdad es una correspondencia es la más natural, desde luego. Quizá sea imposible hallar *algún* filósofo anterior a Kant que *no* mantuviera una teoría de la verdad-correspondencia.

Michael Dummett ha trazado recientemente[3] una distinción entre las perspectivas *no-realistas* (es decir, las que estoy denominando «internalistas») y las *reduccionistas*, con vistas a señalar que los reduccionistas pueden ser realistas metafísicos, es decir, subscriptores

[3] DUMMETT expone sus opiniones en «What is a Theory of Meaning, I, II», en *Truth and Other Enigmas,* Harvard, 1980. Están desarrolladas con más detalle en sus Conferencias William James, dadas en Harvard en 1976.

de la teoría de la verdad-correspondencia. El reduccionismo con respecto a una clase de afirmaciones (por ejemplo, afirmaciones con respecto a eventos mentales), es la concepción que mantiene que las afirmaciones de esa clase son «hechas verdaderas» por hechos que se encuentran fuera de esa clase. Por ejemplo, de acuerdo con cierto tipo de reduccionismo, los hechos acerca de la conducta «hacen verdaderas» las afirmaciones acerca de eventos mentales. Por poner otro ejemplo, la opinión del obispo Berkeley de que todo lo que «realmente existe» son las mentes y sus sensaciones también es *reduccionista,* puesto que sostiene que las oraciones acerca de mesas, sillas y otros «objetos materiales» ordinarios son hechas efectivamente verdaderas por hechos acerca de sensaciones.

Si un punto de vista es reduccionista con respecto a las afirmaciones de un tipo, pero con el único ánimo de insistir en la teoría de la verdad-correspondencia para las oraciones de la clase *reductora,* entonces el punto de vista es realista metafísico en su raíz. Un punto de vista auténticamente no-realista es no-realista en todo su recorrido.

Es frecuente cometer el error de considerar a los filósofos reduccionistas como no-realistas, pero Dummett sin duda está en lo cierto; los primeros discrepan de otros filósofos *en lo que realmente hay,* y no en su concepción de la verdad. Si soslayamos este error, mi afirmación de que es imposible encontrar un filósofo anterior a Kant que *no* fuese realista metafísico, al menos en aquellas afirmaciones que considera *básicas* o irreducibles, parecerá mucho más plausible.

La forma más vetusta de la teoría de la verdad-correspondencia, que perduró aproximadamente 2.000 años, es la que los filósofos antiguos y medievales atribuyeron a Aristóteles. No estoy seguro de que Aristóteles realmente la mantuviese, aunque el lenguaje que emplea así lo sugiere. La denominaré *teoría de la referencia-similitud,* ya que sostiene que la relación que se da entre nuestras representaciones mentales y los objetos externos a los que se refieren es la relación de *similitud* literal.

La teoría empleó, como las teorías modernas, la idea de representación mental. Esta representación, imagen mental del objeto externo, fue llamada *fantasma* por Aristóteles. La relación habida entre el fantasma y el objeto externo, en virtud de la cual el primero representa para la mente el segundo, consiste en que (de acuerdo con Aristóteles) *el fantasma comparte una forma con el objeto externo.* Ya que el fantasma y el objeto externo son similares (comparten la forma), la mente, al tener acceso al fantasma, también accede directamente a la propia forma del objeto[4].

El propio Aristóteles dice que el fantasma *no* comparte con el ob-

[4] Véase *De Anima,* Libro III, Caps. 7 y 8.

jeto propiedades, como la *rojez* (es decir, la rojez en nuestras mentes no es la misma propiedad literal que la *rojez* del objeto), que pueden ser percibidas por un solo sentido, pero sí comparte propiedades como la *longitud o la forma* que pueden ser percibidas por más de un sentido (que son los sensibles comunes, en oposición a los «sensibles propios»).

En el siglo diecisiete, la teoría de la similitud comenzó a restringirse aún más de lo que la había restringido Aristóteles. Así, Locke y Descartes sostuvieron que en el caso de una cualidad «secundaria», como cierto color o cierta textura, sería absurdo suponer que la propiedad de la imagen mental es *literalmente* la misma que la de la cosa física. Locke fue un *corpuscularista,* es decir, un defensor de la teoría atómica de la materia, y pensaba, cual físico moderno, que la rojez sensorial que ofrece mi imagen de un trozo de tela no responde a una propiedad simple de la tela, sino a una propiedad disposicional o facultad *(power)*, a saber, la facultad de dar lugar a sensaciones de este tipo particular (sensaciones que exhiben «rojo subjetivo», en lenguaje psicofísico). Esta capacidad tiene su consecuente explicación, desconocida en los tiempos de Locke, en la particular microestructura del trozo de tela, que le hace absorber y reflejar selectivamente luces de diferente longitud de onda. (Esta *especie* de explicación ya fue ofrecida por Newton.) Si afirmamos que tener esa microestructura es «ser rojo» en el caso del trozo de tela, es obvio que cualquiera que sea la naturaleza de mi rojo subjetivo, el evento que tiene lugar en mi mente (o incluso en mi cerebro) cuando tengo una sensación de rojo objetivo *no* implica que algo sea rojo en mi mente o en mi cerebro. Aquellas propiedades de una cosa física que la convierten en instancia de rojo físico y las propiedades del evento mental que lo convierten en instancia de rojo subjetivo son bastante diferentes. Un trozo de tela roja y una post-imagen roja *no* son literalmente similares. No comparten una Forma.

Sin embargo, Locke estaba dispuesto a salvaguardar la teoría de la referencia-similitud para aquellas propiedades (figura, movimiento, posición) a las que su filosofía corpuscularista le llevó a considerar como básicas e irreducibles. (En realidad, algunos exegetas de Locke discuten esto todavía hoy, pero de hecho Locke afirma que hay una similitud entre la idea y el objeto en el caso de las cualidades primarias, y que no la hay entre la idea de *rojo* o de caliente y la rojez o el calor en el objeto [5]. Y mi lectura de Locke fue la universal, tanto entre sus contemporáneos como en el siglo XVIII.)

---

[5] Véase *Ensayo Sobre el Entendimiento Humano,* Libro III, Cap. VIII.

LA *TOUR DE FORCE* DE BERKELEY

Berkeley descubrió una inoportuna consecuencia de la teoría de la referencia-similitud: implica que nada existe excepto las entidades mentales («los espíritus y sus ideas», es decir, las mentes y sus sensaciones). Por lo general no se aprecia que las premisas de las que partió Berkeley —la teoría de la similitud— no eran algo meramente aprendido de Locke (o leído en Locke), sino que constituían la teoría de la referencia aceptada en épocas anteriores y admitida, en realidad, hasta un siglo más tarde; pero ya hemos subrayado cuán venerable fue tal teoría.

El argumento de Berkeley es muy simple. El argumento filosófico usual contra la teoría de la similitud en el caso de las cualidades secundarias (el argumento de la relatividad de la percepción) es correcto, pero vale también en el caso de las cualidades primarias. La longitud, la figura y el movimiento de un objeto se perciben de forma diferente según los diferentes perceptores o según un mismo perceptor en diferentes ocasiones. Preguntar si una *mesa* tiene la misma longitud que mi imagen de ella o que la imagen que *usted* tiene de ella es formular un interrogante absurdo. Si la mesa mide un metro, y tengo una visión de ella clara y óptima, ¿tengo una *imagen mental de un metro de largo?* Formular la pregunta es ver su sin sentido. Las imágenes mentales no tienen longitud *física*. No pueden compararse con la barra patrón de medida de París. La longitud física y la longitud subjetiva deben ser tan diferentes como la rojez física y la rojez subjetiva.

Formulando la conclusión de Berkeley de otro modo, *nada puede ser similar a una sensación o imagen salvo otra sensación o imagen.* Habida cuenta de esto y dado el supuesto (todavía incuestionado) de que el mecanismo de la referencia es la similitud entre nuestras «ideas» (es decir, nuestras imágenes o «fantasmas») y lo que representan, se desprende inmediatamente que ninguna «idea» (imagen mental) puede representar o referirse a otra cosa que no sea una imagen o una sensación. Sólo podemos concebir, pensar y referirnos a objetos fenoménicos. Y si usted no puede pensar en algo, no puede pensar que existe. El discurso sobre objetos materiales resulta completamente ininteligible salvo que lo consideremos como una derivación del discurso sobre las regularidades de nuestras sensaciones.

La tendencia, en su propio tiempo y también más tarde, a considerar a Berkeley como un loco perverso que, aunque brillante, rayaba en la calumnia, fue provocada por su inaceptable conclusión de que la materia no existe realmente (excepto como una construcción a partir de nuestras sensaciones) y no por alguna peculiaridad de sus premisas. Pero el hecho de que se pueda derivar una conclusión tan inaceptable a partir de la teoría de la similitud produjo una crisis en

la filosofía. Los filósofos que no quisieron seguir a Berkeley en el idealismo subjetivo tuvieron que proponer una descripción diferente de la referencia.

LA TEORIA KANTIANA DEL CONOCIMIENTO Y DE LA VERDAD

Sugiero que la mejor lectura de Kant consiste en considerarlo como el primer autor que propuso lo que he denominado la perspectiva «internalista» o «realista interna» con respecto a la verdad, pese a que Kant en ningún momento afirme explícitamente que lo esté haciendo.

Para empezar, no hay duda de que Kant consideró inaceptables tanto el idealismo subjetivo de Berkeley (así lo afirma explícitamente) como el realismo causal —el punto de vista que pretende que sólo percibimos directamente sensaciones e *inferimos* los objetos materiales mediante algún tipo de problemática inferencia. Según Kant, la opinión según la cual es únicamente una hipótesis muy dudosa que hay una mesa delante de mí mientras escribo estas páginas es una provocación escandalosa.

En segundo lugar, creo que Kant observó con claridad cómo operaba el argumento de Berkeley: cómo dependía de la teoría de la referencia-similitud, y cómo la refutación de esta teoría era un requisito para la del argumento de Berkeley. Atribuyo aquí a Kant un punto de vista que Kant no expresa con estas palabras (en realidad, sólo recientemente se llama «referencia» a la relación entre los signos mentales y aquello que representan, si bien tal relación constituye un antiguo problema). Pero veremos que lo que Kant afirmó tuvo precisamente el efecto de desahuciar la teoría de la referencia-similitud.

Permítaseme sugerir un modo de leer a Kant que puede servirnos de ayuda, aunque es sólo una primera aproximación a una interpretación correcta. Consideremos que Kant aceptaba el punto de vista de Berkeley según el cual el argumento de la relatividad de la percepción se aplica tanto a las cualidades llamadas «primarias» como a las secundarias, pero que elaboró una respuesta diferente de la de Berkeley. Recordemos que la respuesta de Berkeley era descartar la distinción entre cualidades primarias y secundarias y recurrir a lo que Locke había denominado cualidades «simples» de la sensación como entidades básicas a las que podemos referirnos. Recordemos que conforme al tratamiento lockeano de las cualidades secundarias, únicamente podemos concebirlas (en tanto que propiedades del objeto físico) como «facultades», como propiedades —de *naturaleza no especificada*— que permiten que el objeto nos afecte de cierta manera. Decir que algo es rojo, o caliente, o peludo, es decir que es de tal y cual

modo *en relación a nosotros,* y no desde el punto de vista del Ojo de Dios.

Sugiero que (como una primera aproximación) la forma óptima de leer a Kant es como si generalizase a todas las cualidades lo que Locke afirmó con respecto a las cualidades primarias: a las simples, a las primarias y, del mismo modo, a las secundarias (en realidad se distinguen en muy pocos aspectos)[6].

¿Qué se sigue de que *todas las propiedades sean secundarias*? Se sigue que todo lo que decimos sobre un objeto tiene la forma: éste es tal como nos afecta de tal y cual modo. *Nada* de lo que afirmamos acerca de un objeto describe el objeto tal como es «en sí mismo», independientemente de su efecto sobre *nosotros,* sobre seres con nuestra naturaleza racional y con nuestra constitución biológica. Se sigue también que no podemos suponer ningún tipo de similitud («similitude» en el inglés de Locke) entre nuestra idea de un objeto y cualquiera que sea la realidad independiente-de-la-mente responsable de nuestra experiencia del objeto. Nuestras ideas de los objetos no son *copias* de cosas independientes de la mente.

Así es como Kant describe en gran parte la situación. Kant no duda de que hay *alguna* realidad independiente de la mente; para él es virtualmente un postulado de la razón. Alude con diversos términos a los elementos de esta realidad independiente de la mente: cosas-en-sí mismas *(Ding an sich);* objetos nouménicos o *noúmena;* colectivamente, el *mundo nouménico.* Pero no podemos formarnos una concepción real de tales cosas nouménicas; la noción de mundo nouménico es más bien un tipo de límite del pensamiento *(Grenz-Begriff)* que un concepto claro. Esta noción se concibe hoy como un elemento metafísico innecesario en el pensamiento de Kant. (Pero quizá Kant esté en lo cierto, quizá no podamos dejar de pensar que hay, de *algún modo,* una «base» independiente de la mente para nuestra experien-

[6] En los *Prolegómenos,* KANT nos da un resumen de su propio punto de vista, de esta guisa: «Mucho antes de los tiempos de Locke, pero principalmente después de éste, se ha aceptado y concluido que, sin perjuicio de la existencia real de las cosas exteriores, se puede decir de multitud de sus predicados que no pertenecen a las cosas en sí mismas, sino solamente a sus apariencias, y que no tienen existencia propia alguna fuera de nuestra representación. A estos predicados pertenecen el calor, el color, el gusto, etc. Pero si yo, aparte de estas cualidades, aún cuento entre los meros fenómenos, por razones de importancia, las cualidades restantes de los cuerpos que se llaman primarias: la extensión, el lugar y sobre todo el espacio, con todo lo de él dependiente (impenetrabilidad o materialidad, forma, etc.), contra esto no se puede alegar el menor fundamento de inadmisiblidad; y del mismo modo que el que sostiene que el color no es una propiedad que dependa del objeto mismo, sino sólo de la modificación del sentido de la vista, no puede ser, por esto, llamado un idealista, del mismo modo mi doctrina no puede ser llamada idealista, porque yo encuentro que, aún más, *todas las propiedades que constituyen la intuición de un cuerpo pertenecen a su apariencia*» (trad. cast. de Julián Besteiro, Aguilar, Buenos Aires, 1968).

cia, aun cuando las tentativas de hablar de ella nos conduzcan de inmediato al sin sentido.)

Al mismo tiempo, el discurso sobre «objetos empíricos» ordinarios *no* trata acerca de cosas-en-sí mismas, sino acerca de cosas-para-nosotros.

El punto realmente sutil es que Kant considera que estos argumentos se aplican *tanto* a los objetos externos *como* a las sensaciones (objetos del sentido interno). Esto puede parecer extraño. ¿Acaso hay algún problema en que una idea se corresponda o no a una sensación? Pero Kant sigue una pista interesante y profunda.

Supongamos que experimento la sensación *E*, y que la describo afirmando que «*E* es la sensación de rojo», por ejemplo. Si «rojo» significa meramente *como esto*, la afirmación completa únicamente significa «*E* es *como esto*» (señalando a *E*), es decir, *E es como E* —y al fin y al cabo, no he realizado ningún juicio. Como dijo Wittgenstein, uno se degrada virtualmente a un gruñido. Por otra parte, si «rojo» es un genuino *clasificador* y estoy afirmando que esta sensación *E pertenece a la misma clase* que las sensaciones que llamo «rojas» *en otras ocasiones*, entonces mi juicio va más allá de lo inmediatamente dado, más allá de la pura *haecceitas (Thatness)*, e involucra una referencia implícita a otras sensaciones no experimentadas en el instante presente y una referencia implícita al *tiempo* (el cual, de acuerdo con Kant, no es algo nouménico, sino más bien una forma en la que ordenamos las «cosas-para-nosotros»)[7].

¿Son las sensaciones que experimento en diferentes ocasiones, y que clasifico como *sensaciones* de rojo, «realmente» (nouménicamente) similares? Es una pregunta que carece de sentido. Si parecen serlo (por ejemplo, si *recuerdo* las sensaciones precedentes siendo semejantes a ésta, y *anticipo* que las futuras sensaciones que clasificaré de este modo se asemejarán a su vez a ésta, tal como en este momento la recuerdo) entonces son similares-para-mí.

Kant afirma una y otra vez, y de diferentes formas, que los objetos del sentido interno *no* son trascendentalmente reales (nouménicos) sino «trascendentalmente ideales» (cosas-para-nosotros) y que son directamente cognoscibles en el mismo grado en que pueden serlo los denominados objetos «externos». Las sensaciones a las que llamo «rojas» no pueden compararse directamente con objetos *nouménicos* en

---

[7] Al describir aquí el punto de vista de Kant mediante un ejemplo tomado de las *Investigaciones Filosóficas* de Wittgenstein estoy siendo deliberadamente anacrónico. Pero el ejemplo de Wittgenstein tiene profundas raíces kantianas; Hegel, que escribió poco tiempo después de Kant y que conocía perfectamente su doctrina, puso de manifiesto precisamente que cualquier juicio, hasta los relativos a impresiones sensoriales, tiene que ir más allá de lo «dado» para ser siquiera un juicio.

orden a observar si tienen la misma propiedad *nouménica* en un grado mayor al que los objetos a los que llamo «monedas de oro» pueden compararse directamente con objetos nouménicos en orden a observar si tienen la misma propiedad nouménica.

La razón por la que la afirmación «Todas las propiedades son secundarias» constituye sólo una *primera* aproximación al punto de vista de Kant es ésta: la afirmación «Todas las propiedades son secundarias» (es decir, *todas las propiedades son facultades),* sugiere que decir que una silla está hecha de pino, o de cualquier otra cosa, es atribuirle una facultad (la disposición a que a nosotros nos parezca hecha de pino) a un objeto nouménico; decir que la silla es marrón es atribuir una facultad diferente a ese *mismo* objeto nouménico, y así sucesivamente. Según este punto de vista, habría un único objeto nouménico correspondiente a cada objeto de lo que Kant llama «la representación», es decir, un único objeto nouménico correspondiente a cada cosa-para-nosotros. Pero Kant *niega* explícitamente esto. Y éste es el punto donde casi afirma que está abandonando la teoría de la verdad-correspondencia.

En realidad, Kant no *afirma* que está abandonando la teoría de la verdad-correspondencia. Por el contrario, afirma que la verdad es «la correspondencia de un juicio con su objeto». Pero esto es lo que Kant llamó «definición nominal de verdad». En mi opinión, sería un grave error identificarla con lo que el realista metafísico quiere decir con «teoría de la verdad-correspondencia». Para decidir si Kant mantuvo lo que el realista metafísico quiere decir con «teoría de la verdad-correspondencia», hemos de ver si mantuvo una concepción realista acerca de lo que llamó «el objeto» de un juicio empírico.

En opinión de Kant, cualquier juicio acerca de objetos externos o internos (cosas físicas o entidades mentales) afirma que el mundo nouménico, como un todo, es de tal modo que ésta es la descripción que construiría un ser racional (un ser con nuestra naturaleza racional), dada la información de que dispone un ser con nuestros órganos sensoriales (un ser con nuestra naturaleza sensible). En *ese* sentido, el juicio adscribe una facultad. Pero la facultad se adscribe al *mundo nouménico como un todo; no* debemos pensar que ya que en nuestra representación hay sillas, caballos y sensaciones, habrá sillas nouménicas, caballos nouménicos y sensaciones nouménicas que les correspondan. *No hay una correspondencia uno-a-uno entre las cosas-para-nosotros y las cosas en sí mismas.* Kant no sólo renuncia a la idea de similitud entre nuestras ideas y las cosas en sí mismas; renuncia incluso a la idea de isomorfismo abstracto. Y esto significa que en su filosofía no hay una teoría de la verdad-correspondencia.

Entonces, ¿qué es un juicio verdadero? Kant cree que tenemos conocimiento *objetivo:* conocemos leyes de las matemáticas, leyes de la geometría, leyes de la física y muchos enunciados sobre objetos indi-

viduales —objetos empíricos, cosas para nosotros. El uso del término «conocimiento» y el uso del término «objetivo» comporta la afirmación de que *a pesar de todo, hay una noción de verdad.* Pero si no es la correspondencia con la forma en que las cosas son en sí mismas, ¿qué es la verdad?

La única respuesta que se puede extraer de los escritos de Kant, es, como he dicho, ésta: un fragmento de conocimiento (es decir, un «enunciado verdadero») es un enunciado que aceptaría un ser racional, a partir de una cantidad suficiente de experiencia de la clase que los seres con nuestra naturaleza pueden obtener efectivamente. Ni tenemos acceso ni podemos concebir la «verdad» en cualquier otro sentido. *La verdad es bondad última de ajuste.*

LA ALTERNATIVA EMPIRISTA

Pese al punto al que ha llegado nuestro argumento, aún sería posible que un filósofo evitase el abandono de la teoría de la verdad-correspondencia y de la teoría de la referencia-similitud restringiéndolas a *sensaciones e imágenes.* Muchos filósofos continuaron creyendo, incluso después de Kant, que la similitud es el mecanismo por el que somos capaces de referirnos a nuestras propias *sensaciones* (y, aunque esto fue más discutido, a las de los demás) y que éste es el caso principal de referencia, desde un punto de vista epistemológico.

De cara a observar por qué este recurso no resuelve el problema, recordemos que el núcleo del argumento de Berkeley era su aseveración de que nada puede parecerse a una «idea» excepto otra «idea», es decir, que no puede haber semejanza entre lo mental y lo físico. De acuerdo con Berkeley nuestras ideas pueden parecerse a otras entidades mentales, pero no pueden parecerse a la «materia».

Debemos detenernos en esta afirmación y percatarnos de que es falsa, y en un aspecto importante. De hecho, *cualquier cosa es similar a cualquier otra en infinitamente muchos aspectos.* Mi sensación de la máquina de escribir en este instante y la moneda en mi bolsillo son ambas similares en cuanto que algunas de sus propiedades (que la sensación se produzca ahora mismo y que la moneda esté ahora mismo en mi bolsillo) son *efectos de mis acciones pretéritas;* no estaría teniendo la sensación si no me hubiese sentado a escribir a máquina; y la moneda no estaría en mi bolsillo si no la hubiese metido ahí. Tanto la sensación como la moneda existen ambas en el siglo XX. He descrito en castellano tanto la sensación como la moneda, etc. Sólo la ingenuidad y el tiempo limitan el número de semejanzas que uno puede encontrar entre *cualesquiera* dos objetos.

«Similitud» puede tener un significado más restringido en un con-

texto determinado, por supuesto. Pero cuando no especificamos, implícita o explícitamente, la *clase* de similitud en cuestión, preguntar meramente ¿«Son similares A y B»? es formular una pregunta vacía.

De este simple hecho se sigue ya que la idea de que la similitud es el mecanismo privado de referencia debe conducir a un regreso infinito. Supongamos, haciendo uso de un ejemplo debido a Wittgenstein, que alguien está intentando inventar un «lenguaje privado», un lenguaje que se refiera a sus propias sensaciones, tal como le son directamente dadas. Ese alguien dirige su atención a una sensación *X* e introduce un signo *E* con el propósito de aplicarlo exactamente a aquellas entidades que son cualitativamente idénticas a *X*. En efecto, se propone aplicar *E* a todas aquellas entidades que son *similares a X* y sólo a ellas.

Si esto es *todo* lo que pretende y si no especifica con *respecto* a qué tiene algo que ser similar a *X* para caer bajo la clasificación *E*, entonces su propósito es vacío, como ya vimos. Porque cualquier cosa es similar a *X* en algún respecto particular.

Por otra parte, si *especifica* tal respecto, y piensa que una *sensación es E si y solo si es similar a X con respecto a R*, entonces, puesto que es capaz de pensar este pensamiento, puede referirse de antemano a las sensaciones para las que está intentando introducir el término *E*, y a las propiedades relevantes de esas sensaciones. Pero ¿cómo obtuvo *esta* capacidad? Si respondemos: «Dirigiendo su atención hacia otras dos sensaciones *Z, W* y concibiendo el pensamiento de que dos sensaciones son similares con respecto a *R* si y solo si son similares a *Z, W*», entonces nos enredamos en un regreso infinito.

La dificultad de la teoría de la referencia-similitud es la misma que la de la teoría de «la cadena causal del tipo apropiado», que mencionamos con anterioridad. Si afirmo meramente «La palabra "caballo" se refiere a los objetos que tienen la propiedad cuya ocurrencia provoca en mí, en ciertas ocasiones, la proferencia "Hay un caballo delante de mí"», entonces la dificultad radica en que hay demasiadas propiedades de esta índole. Por ejemplo, sea *A-C* («Aparición de Caballo») la propiedad de todas las situaciones perceptivas que provocan la respuesta «Hay un caballo delante de mí» a un hablante competente normal. En este caso, la propiedad *A-C* está presente cuando afirmo «Hay un caballo delante de mí» (aun cuando esté experimentando una ilusión); pero «caballo» no se refiere a situaciones con esa propiedad, sino más bien a ciertos animales. La presencia de un animal con la propiedad de pertenecer a un género natural particular y la presencia de una situación perceptiva con la propiedad *A-C* están *ambas* conectadas con mi proferencia «Hay un caballo delante de mí» mediante cadenas causales. De hecho, la presencia de caballos en la Edad de Piedra está conectada con mi proferencia «Hay un caballo

delante de mí» mediante una cadena causal. Y si había *demasiadas* semejanzas como para que la referencia fuese meramente una cuestión de semejanza, también hay *demasiadas* cadenas causales como para que la referencia sea meramente una cuestión de cadenas causales.

Por otra parte, si digo «La palabra "caballo" se refiere a aquellos objetos que tienen una propiedad conectada con mi emisión, en ciertas ocasiones, de la proferencia "Hay un caballo delante de mí" mediante una *cadena causal del tipo apropiado*», entonces tropiezo con la dificultad de que, si soy capaz de especificar cuál es el tipo de cadena causal, debo ser capaz de referirme de *antemano* a las clases de cosas y propiedades que constituyen ese tipo de cadena causal. Pero ¿cómo obtuve tal capacidad?

De aquí no se concluye que no existan términos que tengan su lógica adscrita por medio de la teoría de la similitud, ni tampoco que no existan términos que se refieren a cosas que estén conectadas con nosotros por determinados tipos de cadenas causales. La conclusión es que ni la similitud ni la conexión causal pueden ser los mecanismos de la referencia únicos o fundamentales.

WITTGENSTEIN SOBRE «SEGUIR UNA REGLA»

Consideremos el ejemplo, que mencioné de pasada, del hombre que trata de especificar el respecto *R* (conforme al cual las sensaciones deben asemejarse a *X* si han de clasificarse correctamente como *E)* diciendo o pensando que las cosas son similares con respecto a *R* sólo en caso de que lo sean precisamente del modo en que son similares *Z, W*. Este intento fracasa, por supuesto, ya que cualesquiera dos cosas *Z, W* son similares de más de una manera (de hecho, de un número infinito de maneras). Intentar especificar una relación de semejanza dando un número finito, aunque amplio, de ejemplos, es como tratar de especificar una función en el dominio de los números naturales dando sus primeros 1.000 (o 1.000.000) valores: hay siempre un número infinito de funciones que coinciden con cualquier conjunto finito de valores de una tabla dada, pero que difieren en valores no inscritos en la tabla.

Esto enlaza con otro punto que Wittgenstein subrayó en las *Investigaciones Filosóficas* y que se mencionó al final del capítulo 1. Cualesquiera signos susceptibles de introspección o representaciones que yo sea capaz de evocar en conexión con un concepto, no puede especificar ni constituir *el contenido* del concepto. Wittgenstein subrayó este punto en una famosa sección en la que se ocupa de «seguir una regla» —por ejemplo, la regla «suma 1». Aun cuando dos especies, en dos mundos posibles (enuncio el argumento en una terminología de lo más anti-wittgensteiniana), tengan los mismos signos mentales en

relación con la fórmula verbal «suma 1» sería posible que sus *prácticas* difiriesen, y es la práctica la que fija la interpretación: los signos no se autointerpretan, como hemos visto. Aun cuando alguien imaginase la relación «*A* es el sucesor de *B*» (es decir $A = B + 1$) tal como lo hacemos nosotros, y estuviese de acuerdo con nosotros en un conjunto de casos amplio pero finito (por ejemplo, en que 2 es el sucesor de 1, 3 es el sucesor de 2, ... 999.978 es el sucesor de 999.977), a pesar de todo, podría tener una interpretación discrepante de «sucesor», que sólo se revelará en algunos casos futuros (aun cuando concuerde con nosotros en su «teoría» —es decir, en lo que *dice* acerca de «sucesor de»— puede tener una interpretación diferente de toda la teoría, como muestra el teorema de Skolem-Löwenheim).

Esto tiene relevancia inmediata tanto para la filosofía de la matemática como para la filosofía del lenguaje. Ante todo, existe el problema del *finitismo:* la práctica humana, actual y potencial, sólo es finitamente prolongable. Aun cuando digamos que podemos, no podemos «seguir contando eternamente». *Si existen posibles ampliaciones divergentes de nuestra práctica, entonces también existen interpretaciones divergentes hasta de la secuencia de los números naturales* —nuestra práctica, nuestras representaciones mentales, etc., no seleccionan un único modelo *estándar* de la secuencia de los números naturales. Nos seduce pensar que lo hace porque pasamos con facilidad desde «podríamos seguir contando» hasta «una máquina ideal podría seguir contando» —(o hasta «una *mente* ideal podría seguir contando»); pero hablar de máquinas (o mentes) ideales es muy diferente a hablar de personas y máquinas *efectivas.* Hablar de lo que una máquina ideal podría hacer es *hablar dentro de* la matemática, y así no puede fijarse la interpretación *de* la matemática.

Del mismo modo, Wittgenstein sostiene que hablar de «similitud» y de «la misma sensación» o de «la misma experiencia» es hablar *dentro* de la teoría psicológica; así no puede fijarse la interpretación *de* la teoría psicológica. *Esta,* la interpretación de la terminología y de la teoría psicológica, se fija mediante nuestra práctica efectiva, mediante nuestros estándares usuales de corrección e incorrección.

En *Ways of Worldmaking*[8], Nelson Goodman recalca un punto íntimamente relacionado: es fútil esforzarse en tener una noción de lo que «realmente son» los hechos perceptivos, independientemente de cómo los conceptualizamos, de las descripciones que de ellos demos y que nos parezcan correctas. Así, después de hablar de un hallazgo de Kolers, psicólogo que se percató de que un número desproporcionado de físicos o ingenieros son absolutamente incapaces de ob-

[8] Publicado por Hackett, 1978.

servar el movimiento aparente, esto es, el «movimiento» producido por luces que lanzan destellos sucesivamente desde posiciones distintas, Goodman comenta (p. 92):

> Con todo, si un observador nos informa de que ve dos destellos distintos incluso a distancias e intervalos tan cortos que la mayoría de los observadores ve un punto móvil, quizá quiera decir que ve dos destellos así como nosotros podríamos decir que vemos un enjambre pululante de moléculas cuando miramos una silla, o así como podríamos decir que vemos la superficie circular de una mesa aun cuando la observásemos desde un ángulo oblicuo. Ya que un observador puede convertirse en un experto a la hora de distinguir el movimiento aparente del real, podría considerar la apariencia de movimiento como señal de que hay dos destellos, así como nosotros consideramos la apariencia elíptica de la mesa como señal de que es circular: y en ambos casos las señales pueden ser, o llegar a ser, tan transparentes, que veamos a través de ellas los acontecimientos físicos y los objetos. Cuando un observador determina visualmente que lo que está ante él es aquello que reconocemos que está ante él, difícilmente podemos achacarle un error en su percepción visual. ¿Diremos, más bien, que comprende mal la instrucción, la cual es, presumiblemente, que diga lo que ve? Entonces, ¿cómo podemos reconstruir nuestra pregunta de cara a prevenir tal «malcomprensión» sin perjudicar el resultado? Pedirle que no haga uso de experiencias previas y que eluda toda conceptualización es condenarlo al silencio; porque para hablar debe usar palabras.

APREHENSION DE «FORMAS» Y ASOCIACION EMPIRICA

Un platónico o neoplatónico a la antigua usanza habría afrontado este problema de una manera mucho más simple. Tal filósofo habría dicho que cuando prestamos atención a una sensación particular también percibimos un Universal o una Forma, es decir, que la mente tiene la capacidad de aprehender propiedades en sí mismas, y no sólo de prestar atención a instancias de esas propiedades. Tal filósofo diría que es el nominalismo de Wittgenstein y Goodman, su negativa a mantener relaciones con las Formas y con la aprehensión directa de Formas, lo que les hace ver problemática la teoría de la similitud.

Aun cuando postular sin más una misteriosa facultad de «aprender Formas» difícilmente puede ser una solución, podría parecer que disponemos de una facultad análoga. Las propiedades de las cosas toman parte en las *explicaciones causales;* cuando experimento una sensación y ésta obtiene la respuesta «Esta es la sensación de rojo», mi respuesta está causada, en parte, por el hecho de que la sensación tiene una propiedad. Es cierto que algunos filósofos son tan nominalistas que negarían por completo la existencia de entidades tales como las «propiedades»; pero la propia ciencia no puede vacilar en hablar libremente de propiedades. Cuando el personaje de Wittgenstein (el hombre que quiere inventar su lenguaje privado) señala *X* y dice «*E*», ¿no podemos decir que lo que causó la respuesta «*E*» fue una interac-

ción causal que involucra cierta *propiedad,* y que esta propiedad (sea cual sea) constituye la similitud relevante que otras sensaciones deben mantener con *X* para ser clasificadas correctamente como *«E»?*

La observación de que es perfecta y científicamente legítimo hablar de propiedades es correcta: pero esto no ayuda a la rehabilitación del platonismo. Interactuamos con propiedades sólo mediante la interacción con sus *instancias;* y estas instancias son siempre instancias de *muchas* propiedades al mismo tiempo. No hay tal cosa como una interacción precisa con una propiedad «en sí misma». El discurso sobre propiedades causalmente asociadas con una sensación no puede realizar la labor que la idea de Forma (única) de la sensación realizaba en la filosofía platónica.

Detallando: cuando experimento la sensación de azul experimento además de una sensación de *azul* una sensación con la compleja propiedad de ser de tal suerte que debo clasificarla en ese caso bajo esa particular etiqueta. Atendiendo meramente a esta sensación no aprehendemos *una* de estas propiedades. Discernir la propiedad asociada con mi sensación o con la etiqueta verbal precisamente en *uno* de estos modos conduce a nuestro viejo amigo, el problema de la cadena causal del tipo apropiado.

Para comprenderlo, observemos en primer lugar que cuando mi experiencia perceptiva total provoca la respuesta «Estoy experimentando la sensación de azul», no siempre estoy en lo *cierto.* Yo mismo he tenido la experiencia de referirme dos o tres veces a «el hombre del suéter azul», antes de que alguien me indicase que el suéter era *verde.* No quiero decir que el suéter *pareciera* azul; en cuanto la otra persona habló, me di cuenta de que estaba describiendo erróneamente el suéter. (No tengo a menudo la ocasión de decir «Estoy experimentando la sensación de azul», pero en caso de haberla tenido, lo habría dicho dos o tres veces antes de que alguien —preguntándose por qué tengo la sensación de azul cuando estoy mirando algo que es obviamente verde— me expresase sus dudas, con lo cual me retractaría de mi informe fenoménico previo.) Solamente esto muestra ya que la propiedad de provocar el informe «Estoy experimentando la sensación de azul» *no es la misma* que la propiedad de ser una sensación de azul o una sensación de la cualidad relevante que se nos antoje.

Con frecuencia los filósofos denominan a este caso *«lapsus linguae».* Me parece una terminología desafortunada. La palabra «verde» podría haber estado en mis labios y podría haberme sorprendido *diciendo* «azul», decepcionantemente. Esto sí habría sido un *lapsus linguae.* Pero en el caso que describí anteriormente, ni siquiera reparé en que mi descripción era errónea hasta que alguien puso en duda mi informe (si no lo hubiera hecho, quizá nunca me hubiera dado cuenta).

Otra explicación que se sugiere es que cuando dije «azul» *quise*

decir verde. Pero ya deberíamos tener claro que cuando decimos cosas no damos vueltas en torno a los «significados» de las cosas, como si nuestra mente contuviese significados. Afirmar que quise decir «verde» es tan sólo decir que acepté instantáneamente la corrección (y que me extrañé al darme cuenta de lo que había dicho). Esto es sólo repetir lo que sucedió y no explicarlo.

Sea cual sea la explicación (quizá algún desliz en la unidad de procesamiento verbal de mi cerebro), lo importante es que, así como la propiedad *A-C,* descrita pocas páginas atrás, provocará el informe «Hay un caballo delante de mí» aun cuando en el entorno no haya presente ningún caballo, del mismo modo, hay una compleja propiedad de mi aparato mental, que provocará «Estoy experimentando la sensación de azul» aun cuando no la esté experimentando (o, de cualquier modo, aun cuando lo niegue si se me pregunta). Ningún mecanismo de asociación empírica es perfecto. Si decidimos estipular que estoy experimentando la sensación de azul cuando experimento una sensación que *provoca* aquel informe (o que lo provoca siempre que éste no me parezca «erróneo» al reexaminarlo), entonces, según la teoría psicológica popular —y a lo mejor también según la científica— podría haber ocasiones en las que, de acuerdo con este *criterio,* será verdadero que estoy experimentando la sensación de azul pese a que, debido a diversas razones, la sensación no tenga la cualidad de azul. Por otra parte, tal como expuso Wittgenstein, según este criterio, *cualquier cosa que me parezca correcta pasa a ser correcta* —es decir, se habrá abandonado la distinción entre elaborar un informe de mi sensación realmente correcto y elaborar un informe de mi sensación que me parezca correcto. Quizá *debiéramos* abandonarla, o al menos atenuarla; quizá —como Goodman parece sugerir— no tenga sentido que uno se pregunte si está realmente experimentando el tipo de sensación que piensa estar experimentando, salvo casos especiales —como el caso en el que uno se retractaría de su informe si alguien le expresase sus dudas; pero un realista metafísico no puede jugar la baza de abandonar esta distinción, ya que lo que caracteriza la postura del realismo metafísico es la tajante distinción entre lo que realmente es y lo que uno juzga que es.

¿SE PUEDE ESTAR SIEMPRE EQUIVOCADO CON RESPECTO A LA CUALIDAD DE LAS SENSACIONES QUE SE TUVIERON EN EL PASADO?

Otra manera de sacar a relucir en qué andamos metidos es examinar la pregunta «¿Se puede estar siempre equivocado con respecto a la cualidad de las sensaciones que se tuvieron en el pasado?». Según

la teoría de la similitud, la respuesta es claramente «sí», porque de acuerdo con esta teoría, mis sensaciones previas o bien son o bien no son similares a las sensaciones que *ahora* describo mediante diversas etiquetas verbales, como «sensación de rojo», «dolor», etc., y si lo son o no es una cuestión completamente distinta de si las clasifiqué *entonces* bajo esas mismas etiquetas verbales. A lo mejor el mundo es de tal forma que lo que llamamos «sensación de rojo» en un minuto par a partir del comienzo de la era cristiana es cualitativamente similar a la que llamamos «sensación de verde» en un minuto impar, pero nuestra memoria siempre nos engaña, de tal manera que nunca lo notamos. La sensación que clasifiqué bajo la etiqueta verbal «sensación de rojo» hace un minuto *no* sería similar a la sensación que ahora clasifico bajo la misma etiqueta.

Sin embargo, hay algo extraño en esta pretendida posibilidad. En primer lugar, el sentido en el cual «nunca lo notaría» está muy acentuado: si considero que mis sensaciones de rojo son un signo fiable de las diversas ocurrencias físicas correlacionadas, tales como el fuego, la señal de «stop», etc., entonces tendré éxito en todas mis acciones. La clase «errónea» de similitud (la clase que amontona juntas las sensaciones a las que *llamo* rojas, a pesar de que en realidad no tienen todas la misma «cualidad») sería aquella de la que hice *mejor* uso en relación con mis actividades de resolución de problemas. Pero entonces ¿es realmente la clase *errónea* de similitud?

Si no suponemos que la noción de similitud se autointerpreta, este caso podría redescribirse como un caso en el que la relación que el observador externo —que nos está hablando del caso— llama «similitud» simplemente difiere de la relación que *nosotros* llamamos «similitud». Si adoptamos este punto de vista, fracasará la hipótesis de que estemos realmente equivocados con respecto a nuestras sensaciones pasadas: desde un punto de vista *internalista* no existe ninguna noción inteligible de similitud de sensaciones en diferentes tiempos, aparte de nuestros estándares de aceptabilidad racional.

OTRA VEZ LA TEORIA DE LA VERDAD-CORRESPONDENCIA

El lector puede haberse convencido ya de que la teoría de la referencia-similitud está definitivamente muerta. Pero ¿por qué hemos de concluir que debe abandonarse la teoría de la verdad-correspondencia? Aun cuando la noción de «similitud» entre nuestros conceptos y sus referentes no solucione el problema, ¿no podría haber algún tipo de isomorfismo abstracto, o sin ser literalmente isomorfismo, algún tipo de proyección de conceptos sobre cosas en el

mundo (independiente de la mente)? ¿No podría definirse la verdad en términos de tal isomorfismo o «proyección»?

Esta sugerencia tropieza con la dificultad, no de que no existan correspondencias entre palabras o conceptos y otras entidades, sino de que existen *demasiadas* correspondencias. Para seleccionar exactamente *una* correspondencia entre las palabras o los signos mentales y las cosas independientes de la mente, deberíamos tener de antemano acceso referencial a las cosas independientes de la mente. No podemos seleccionar una correspondencia entre dos cosas tomando como fulcro a *una* de ellas (o haciendo cualquier otra cosa sólo a una de ellas); usted no puede seleccionar una correspondencia entre conceptos y los supuestos objetos nouménicos sin tener acceso a estos últimos.

Una vía para entender esto es la que sigue. Las teorías incompatibles pueden ser efectivamente intertraducibles en algunas ocasiones. Por ejemplo, si la física newtoniana fuese verdadera, entonces cada evento físico determinado podría describirse de dos formas: en términos de partículas actuando a distancia, a través de un espacio vacío (así es como Newton describió la acción gravitatoria), o en términos de partículas que actúan sobre campos, que a su vez actúan sobre otros campos (o sobre otras partes del mismo campo), los cuales, finalmente actúan «localmente» sobre otras partículas. Por ejemplo, las ecuaciones de Maxwell, que describen la conducta del campo electromagnético, son matemáticamente equivalentes a una teoría que contenga solo fuerzas de acción-a-distancia entre partículas, atrayéndose y repeliéndose de acuerdo con la ley de la proporción inversa de los cuadrados, desplazándose no instantáneamente, sino más bien a la velocidad de la luz («potenciales retardados»). La teoría del campo de Maxwell y la teoría de los potenciales retardados son incompatibles desde un punto de vista metafísico, ya que o bien hay o bien no hay agentes causales (los «campos») que medien la acción de partículas separadas entre sí (diría un realista). Pero las dos teorías son matemáticamente intertraducibles. Así que, si hay una «correspondencia» hacia cosas nouménicas que haga verdadera a una de estas teorías, entonces se puede definir otra correspondencia que haga lo propio con la otra. Si se considera que todo lo que hace verdadera a una teoría es una correspondencia abstracta (no importa cuál), entonces las teorías incompatibles pueden ser verdaderas.

Para un internalista, esto no es objetable: ¿por qué no han de existir a veces esquemas conceptuales igualmente coherentes, pero incompatibles, que se ajusten igualmente bien con nuestras creencias experienciales? Si la verdad no es una (única) correspondencia, entonces se abre la posibilidad de cierto pluralismo. Pero la motivación del realista metafísico es salvar la noción del Punto de Vista del Ojo de Dios, esto es, La Teoría Verdadera.

No sólo puede existir una correspondencia entre objetos y (lo que nosotros consideramos) teorías incompatibles (es decir, *los mismos objetos* pueden ser lo que los lógicos llaman «un modelo» para teorías incompatibles), sino que aun cuando fijemos la teoría y los objetos, *hay* una cantidad infinita (si el número de objetos es infinito) de formas *diferentes* en las que pueden usarse los mismos objetos para construir un modelo para una teoría dada. Esto es simplemente enunciar en lenguaje matemático el hecho intuitivo de que para seleccionar una correspondencia entre dos dominios necesitamos tener acceso independiente a ambos dominios.

Nos encontramos ante el óbito de una teoría que resistió alrededor de dos siglos. El que persistiera tanto y de tantas formas, a pesar de sus obscuridades y contradicciones internas, presentes desde un principio, da fe de la naturalidad y de la fuerza del deseo del punto de vista del Ojo Divino. Kant, que fue el primero en enseñarnos la insatisfacibilidad de tal deseo, pensó que, no obstante, formaba parte de nuestra propia naturaleza racional (sugirió sublimar este «impulso totalizador» en el proyecto de realizar el mayor bien en el mundo, reconciliando los órdenes moral y empírico en un sistema de instituciones sociales y relaciones individuales perfeccionado). La continua presencia de este impulso natural pero insatisfacible quizá sea el motivo de fondo de los falsos monismos y dualismos que proliferan en nuestra cultura; sea como sea, nos hemos quedado sin el punto de vista del Ojo Divino.

4. MENTE Y CUERPO

PARALELISMO, INTERACCIONISMO E IDENTIDAD

En el siglo diecisiete, grandes filósofos como Descartes, Spinoza y Leibniz advirtieron que la relación entre la mente y el cuerpo material constituía un serio problema. No hay duda de que, en cierta medida, ya lo había sido para Platón y para todos los filósofos que vinieron después, pero con el nacimiento de la nueva *física* esa relación se convirtió en algo mucho más problemático. Los hombres del siglo diecisiete se convencieron de que el mundo físico *estaba causalmente cerrado.* La mejor forma de expresar la clausura causal del mundo es con los términos de la física newtoniana: ningún cuerpo se mueve excepto como resultado de la acción de alguna *fuerza.* Las fuerzas pueden describirse exhaustivamente mediante números: tres números son suficientes para determinar la dirección de cualquier fuerza, y un número basta para describir su magnitud. La aceleración producida por una fuerza tiene exactamente la misma dirección que esa fuerza; además, la magnitud de la aceleración puede deducirse de la masa del cuerpo y de la magnitud de la fuerza, de acuerdo con la primera ley de Newton $F = m \times a$. Cuando sobre un cuerpo actúa más de una fuerza, la fuerza resultante puede calcularse mediante la ley del paralelogramo.

Es importante reconocer lo lejos que se encuentra esta física del pensamiento esencialmente *cualitativo* de la Edad Media, principalmente en el énfasis que pone en los números y en la precisión de los algoritmos de cómputo. En el pensamiento medieval, (casi) cualquier cosa podía ejercer influencia sobre cualquier otra. (Nuestra palabra «influencia» es un vestigio del modo de pensar del medievo. Las fuerzas del mal se concebían como malignos espíritus que ejercían una influencia —*questa influenza,* en italiano— sobre el aire, el cual influía a su vez sobre los que padecían una enfermedad.) Teniendo en cuenta estos supuestos, no es tan sorprendente que la mente pudiera ejercer «influencia» sobre el cuerpo.

En el tiempo de los filósofos mencionados, el modo matemático de pensar comenzaba a ver la luz y a arrinconar al pensamiento medieval. Pese a que no se desarrolló completamente hasta Newton, Descartes ya usó el paralelogramo de fuerzas en determinados casos; y Leonardo da Vinci también lo usó, aunque en casos todavía más sencillos. Estos pensadores observaron que podía hacerse física de una manera que se parece bastante a la actual. Observaron que la física

trata de la fuerza y del movimiento, y, en consecuencia, rechazaron el estilo cualitativo de explicación. Mejor dicho, se convencieron de que el mundo mecánico poseía una lógica propia, o un «programa», como nosotros diríamos, y que seguía el programa salvo que algo lo alterase.

A estos pensadores les parecía que con los eventos mentales podían ocurrir una de estas dos cosas: (1) Podían ser *paralelos* a los eventos físicos, por ejemplo, a eventos que acaecen en el cerebro. El modelo es un par de relojes sincronizados: el cuerpo es un reloj al que se le ha dado cuerda y que funciona, feliz o infelizmente, hasta su muerte; del mismo modo, el mundo físico funciona feliz o infelizmente desde la creación hasta el Juicio Final (o hasta el colapso gravitatorio, según la versión más moderna). También los eventos mentales ocurren, feliz o infelizmente y, de algún modo, quizá gracias a la providencia divina, se ha dispuesto que ese evento mental *B* ocurra siempre y cuando esté ocurriendo la sensación *S*. (2) Podían *interactuar* con los eventos físicos. Los eventos mentales podrían estar causando los eventos cerebrales, y viceversa.

La más célebre formulación cartesiana del punto de vista interaccionista, la sugerencia de que la mente puede influir en la materia cuando la materia es sumamente etérea (de hecho empuja a la materia en la glándula pineal) no era una especulación tan disparatada como parece, sino más bien un residuo de un conjunto de doctrinas medievales [1]. El pensamiento medieval consideraba que la mente actuaba sobre el «espíritu», el cual actuaba a su vez sobre la «materia»; pero el espíritu no se concebía como algo *totalmente* inmaterial. El «espíritu» era precisamente el tipo de sustancia intermedia que los filósofos medievales se inclinaron a postular, inducidos por su tendencia a introducir entes intermedios entre cualesquiera dos términos adyacentes en la serie de géneros del ser. El espíritu se concebía como un gas con muy poca presión. Tan pronto como se niega la competencia explicativa del «espíritu» y se concibe a la mente como algo totalmente inmaterial empieza a resultar extraño que ésta pueda empujar a la materia, ni siquiera en la glándula pineal, donde se supone que la materia es sumamente etérea.

La versión más ingenua de la perspectiva interaccionista concibe a la mente como una especie de fantasma, capaz de habitar diferentes cuerpos (pero sin cambiar su forma de pensar, sentir, recordar, y de manifestar una personalidad, a juzgar por el torrente de libros populares sobre la reencarnación y sobre el «recuerdo de vidas previas») e incluso capaz de existir sin un cuerpo (y de continuar pensando, sin-

---

[1] Para una descripción de la concepción medieval, v. *The Discarded Image*, de C. S. LEWIS, especialmente el capítulo VIII, Sección F, Cambridge, 1984.

tiendo, recordando y manifestando su personalidad). Esta versión, que equivale a poco más que a un relato supersticioso, es vulnerable a la objeción de que existe enorme evidencia (ya conocida en el siglo diecisiete) de que las funciones del pensamiento, sentimiento y memoria tienen que ver con el cerebro de una forma esencial. En realidad, esta versión no explica porqué tenemos *cerebros* tan complejos, si todo lo que se necesita es un pequeño «volante», que incluso puede ser de menor tamaño que el cerebro humano.

Para eludir estas objeciones científicas, filósofos interaccionistas tan sofisticados como Descartes sostuvieron que la mente y el cerebro constituían una *unidad esencial.* De algún modo, es la unidad mente-cerebro la que piensa, siente, recuerda y manifiesta una personalidad. Esto quiere decir que por lo general llamamos «mente» a algo que no es sólo la mente, sino que es más bien la unidad mente-cerebro. No obstante, no está claro lo que esta doctrina significa, lo que significa decir que algo puede constar de dos sustancias tan distintas como se supone que son la mente y el cerebro y aún así ser una esencial unidad.

La alternativa paralelista también resulta bastante extraña. ¿Qué es lo que hace que el evento mental acompañe al cerebral? Spinoza, un osado filósofo del siglo diecisiete, sugirió que los eventos mentales son en realidad *idénticos* a los eventos cerebrales y a otros eventos físicos. La sugerencia, en su formulación contemporánea, afirma que el que yo sienta dolor en una ocasión determinada puede constituir el mismo evento que el que mi cerebro se halle en algún estado *B* en esa ocasión. (También expondré esta concepción afirmando que las propiedades de sentir un dolor particular y de hallarse en el estado cerebral *B* son idénticas. Prefiero «expresarlo» así, puesto que considero que disponemos de una teoría lógica de *propiedades* en mayor medida que de una de eventos, aunque creo que ambos modos de expresión son correctos). La idea, en esta terminología, es que la persona tiene una propiedad, estar experimentando la sensación *Q,* que puede ser la *misma* propiedad que la propiedad de hallarse en el estado cerebral *B*. Bajo esta formulación, presentada ya por Diderot en el siglo dieciocho, la propuesta se convirtió en la «corriente en boga» en los años cuarenta y cincuenta de este siglo. El materialismo y la teoría de la identidad se tomaban en serio por primera vez, hasta el punto de que comenzó a fomentarse la idea de que una concepción parecida a la de Spinoza (o la concepción de Spinoza *menos* sus elaborados embellecimientos teológicos y metafísicos) era correcta: tratamos en realidad con un mundo, y el hecho de que sólo cuando hemos logrado desarrollar una cantidad considerable de ciencia hemos llegado a saber que los estados de sentir dolor, oír sonidos y experimentar sensaciones visuales, etc., son en realidad estados cerebrales, no significa que no puedan serlo.

Fueron varios los autores que propusieron la primera formulación contemporánea de esta teoría de la identidad, siendo uno de los más conocidos el filósofo australiano J. J. C. Smart. En un principio se sugirió que una sensación, por ejemplo, la sensación particular de azul, es idéntica a cierto estado neurofisiológico. Creo que fui yo quien sugerí por primera vez el punto de vista denominado *funcionalismo*[2], que es una variante de la anterior afirmación. Esta concepción admite la existencia de una identidad en el lugar señalado, pero niega que en el otro término de la identidad figure el tipo de propiedad que Smart señala. De acuerdo con el funcionalista, el cerebro tiene propiedades que, hasta cierto punto, son *no-físicas.*

Ahora bien, ¿qué quiero decir al afirmar que el *cerebro* tiene propiedades *no-físicas?* Quiero decir que posee propiedades que son *definibles en términos que no aluden a la física o a la química cerebral.* Si nos parece extraño que un sistema físico tenga propiedades que no son físicas, prestemos atención a un ordenador. Un ordenador tiene muchas propiedades físicas. Tiene cierto peso, por ejemplo; tiene también cierto número de circuitos o *chips,* o cualquier otra cosa. Tiene propiedades económicas, tales como tener cierto precio; y propiedades funcionales, tales como tener cierto programa. Pues bien, este último tipo de propiedad es no-física, *en el sentido de que puede ser llevada a cabo por un sistema, no importa cuál pueda ser su composición metafísica u ontológica, si es que la tiene.* Un espíritu incorpóreo podría presentar cierto programa, un cerebro podría presentar cierto programa, una máquina podría presentar cierto programa, pero la organización funcional de los tres (el espíritu incorpóreo, el cerebro y la máquina) podría ser exactamente la misma aun cuando su materia, su sustancia, fuese completamente diferente.

Las propiedades psicológicas presentan la misma característica; la misma propiedad psicológica (por ejemplo, estar iracundo) puede ser una propiedad perteneciente a miembros de miles de especies diferentes, que pueden poseer una composición física o química completamente distinta (algunas especies podrían ser extraterrestres; y quizá los robots muestren ira algún día). La sugerencia del funcionalista es que la teoría «monista» más plausible que se puede defender en el siglo XX, la teoría que evita tratar a la mente y a la materia como dos tipos separados de sustancias o como dos reinos separados de propiedades, es la que identifica propiedades psicológicas y propiedades funcionales.

---

[2] El libro de NED BLOCK *Readings in Phylosophy of Psychology,* Harvard, 1980, contiene una excelente colección de artículos sobre el funcionalismo. Mis propios escritos sobre el funcionalismo están reimpresos en los capítulos 14 al 22 de mi libro *Mind, Language and Reality, Philosophical Papers,* Vol. 2, Cambridge, 1975.

Todavía hoy me inclino a pensar que esa teoría es correcta; o al menos que constituye una descripción *naturalista* correcta de la relación mente-cuerpo. Las descripciones «mentalistas» de esta relación también son correctas, aunque no son reducibles a la imagen del mundo que llamamos «Naturaleza» (en realidad, las nociones de «racionalidad», «verdad» y «referencia» *pertenecen* a esta versión mentalista). Más adelante diré algo sobre estas últimas (capítulo 6). Este hecho no me desanima: como ha puesto de relieve Nelson Goodman, uno de los atractivos del no-realismo es que permite la posibilidad de versiones correctas del mundo alternativas. A pesar de todo, me atrae la idea de que *una* de las versiones correctas del mundo es la naturalista, en la que las formas-de-pensamiento, imágenes, sensaciones, etc. *son* acontecimientos físicos funcionalmente caracterizados; lo que deseo discutir aquí es una dificultad de la teoría funcionalista, que se me planteó hace ya algunos años: la teoría funcionalista halla dificultades con el carácter cualitativo de las sensaciones. Cuando se piensa en estados psicológicos puros y relativamente abstractos, por ejemplo, en lo que he denominado creencias «puestas entre paréntesis», esto es, un pensamiento considerado sólo en su contenido «nocional», o en estados emocionales tan difusos como estar celoso o iracundo, su identificación con estados funcionales de todo el sistema parece sumamente plausible; pero cuando se piensa en experimentar una cualidad dada, por ejemplo, experimentar una tonalidad particular de azul, la identificación ya no es plausible.

Durante muchos años he utilizado en mis clases una variante del famoso ejemplo del «espectro invertido». El ejemplo del espectro invertido (que aparece en los escritos de Locke[3]), atañe a un tipo que ve las cosas de tal modo que el azul le parece rojo y el rojo azul (o de tal modo que sus colores subjetivos parecen estar en negativo y no en positivo). La primera reacción al oír un caso así podría ser la siguiente: «Pobre tipo, es digno de lástima». Pero ¿cómo podría saberlo algún otro? Cuando ve algo azul le parece rojo, pero desde la infancia se le ha enseñado a llamar «azul» a ese color, de modo que si se le pregunta cuál es el color del objeto, contestaría que azul. De forma que nadie lo sabría nunca.

Mi variación del ejemplo fue la siguiente: imagine usted que su espectro se invierte en un momento particular de su vida, *pero que recuerda cómo era antes de que eso ocurriera.* La «verificación» no plantea ningún problema epistemológico. Usted se levanta una mañana y el cielo le parece de color rojo, y su jersey rojo parece haberse vuelto azul, además de que todas las caras tienen un color espantoso,

[3] *Ensayo sobre el Entendimiento Humano,* libro II, Cap. 32, Sec. 14.

como en negativo: «¡Por Dios!». Ahora bien, a lo mejor puede aprender a variar su forma de hablar, llamando «azules» a las cosas que le parecen rojas, y quizá podría adquirir la pericia suficiente para ofrecer la respuesta «normal» cuando alguien le preguntase cuál es el color de su jersey. Pero permítanos imaginar su lamento nocturno: «¡Oh, ojalá vuelva a ver los colores como los veía cuando era niño! Ya no los veo como solía verlos».

Da la impresión de que en este caso uno puede saber lo que debe haber sucedido. Se deben haber cruzado algunos «cables» en el cerebro. Los *inputs* de luz azul que solían dirigirse hacia un mecanismo cerebral se dirigen ahora hacia otro, llegando al primero los *inputs* de luz roja. En otras palabras, algo ha trastocado de forma súbita las *materializaciones,* esto es, los estados *físicos.* El estado *físico* que en un principio desempeñaba el *rol* funcional de señalar la presencia de azul «objetivo» en el entorno, señala ahora la presencia de rojo «objetivo» en éste.

Supongamos ahora que adoptamos la siguiente teoría «funcionalista» con respecto al color subjetivo: «una sensación es la sensación de azul (es decir, tienen el caracter *cualitativo* que describo de esa forma *en este momento*) sólo en el caso de que la sensación (o el correspondiente acontecimiento físico en el cerebro) desempeñe el papel de señalar la presencia de azul objetivo en el entorno». Aunque esta teoría capta un sentido de la locución «sensación de azul», no es éste el sentido «cualitativo» deseado. Si el rol funcional fuese *idéntico* al carácter cualitativo entonces no podría afirmarse que la cualidad de la sensación ha sufrido un cambio. (Si esto no queda del todo claro, imaginemos que una vez ocurrida la inversión del espectro, y después de haber aprendido a compensarla lingüísticamente, usted experimenta un ataque de amnesia que le borra por completo de la memoria cómo solía ver los colores. En este caso, podría parecer que la sensación a la que usted llama *ahora* «sensación de azul» desempeña casi el mismo rol funcional que la sensación que usted solía llamar «sensación de azul», a pesar de que tiene un carácter distinto.) Mas la cualidad *ha* cambiado. En *este* caso, la cualidad no parece ser un estado *funcional.*

Si estos casos son realmente posibles, me parece que la baza más plausible que puede jugar un funcionalista es decir: «Sí, pero el "carácter cualitativo" es precisamente la materialización física» y a continuación afirmar que para este tipo especial de propiedad psicológica, para las *cualidades,* la formulación correcta de la teoría de la identidad es la más antigua. Si el lector(a) tiene inclinaciones honestamente materialistas, probablemente piense que la propiedad de tener una sensación es una propiedad del cerebro. Los lectores que no las tengan, probablemente piensen que esa propiedad está *correlacionada* con un estado del cerebro. Y, probablemente, la mayoría de las personas sos-

tienen una de estas dos opiniones: la que afirma que los estados sensoriales están correlacionados con estados del cerebro o la que afirma que los estados sensoriales son idénticos a los estados del cerebro. Como sucede bastante a menudo, el debate cobra esta forma una y otra vez. He aquí como discurre habitualmente la discusión: dado que *B* está correlacionado con *Q*, ¿es *B* efectivamente idéntico a *Q*? Sabemos que este estado sensorial corre paralelo a este estado del cerebro: ¿es el primero idéntico al segundo o no lo es? A medida que la discusión va discurriendo por estos derroteros, el concepto de *correlación* va pareciendo menos problemático. La *correlación* no es una cuestión (muy) debatida, pues *todo el mundo sabe* que hay al menos una correlación. La *identidad* sí que se discute, porque es ahí donde radica el problema, claro está. Frente a esto, voy a intentar demostrar que hasta la *correlación es problemática, y no en el sentido de que exista evidencia a favor de la no-correlación,* sino en el sentido epistemológico de que *aunque* exista una correlación, nunca podemos averiguar cuál es ésta. El problema no radicará en asumir el materialismo, sino en el hecho de que creamos que hay al menos una correlación.

LA TEORIA DE LA IDENTIDAD Y EL *A PRIORI*

Fue el cambio de clima epistemológico el que hizo posible el resurgimiento del interés hacia la teoría de la identidad y hacia otras teorías «monistas»; y aunque este resurgimiento no se dio en un principio (esto es, no se dio ni con Smart ni con los anteriores teóricos de la identidad), empezó a notarse en los comienzos de la década de los 60. La teoría de la identidad no se tomó en serio antes de esta década porque los filósofos «estaban seguros» de que era falsa, y estaban convencidos de que sabían que era falsa *a priori,* no sobre la base de la evidencia empírica (¿qué tipo de evidencia empírica podría mostrar que un estado sensorial no puede ser un estado cerebral?). Basta pensar sobre ello para comprender *a priori que no tiene sentido decir* que un estado sensorial es un estado cerebral, del mismo modo que no lo tiene decir que el número tres es azul. Anteriormente a 1950 ó 1960, muchas personas estaban convencidas de que sabían, *sin más,* que los estados sensoriales no podían ser estados físicos. Otras estaban convencidas de que sabían que las primeras estaban equivocadas. Sin embargo, era imposible proporcionar argumento alguno. La mayoría diría: «Mire usted, no podemos probarle que es imposible que un estado sensorial sea un estado neurofisiológico, ni que cada número tiene un sucesor, ni que el número tres no es azul, pero esto son cosas que sabemos sin más, son *verdades de razón.* Sabemos que es un sinsentido o una imposibilidad que un estado sensorial sea un estado neurofisiológico con tanta claridad como sabemos cualquier otra

cosa». La mayoría estaba convencida de que los estados sensoriales no podían ser estados cerebrales, y una minoría estaba convencida de que la mayoría estaba equivocada. Cada parte *sabía a priori* que la otra estaba en un error. Ante este colapso en el debate no había ninguna posibilidad realmente significativa de jugar una baza o de ofrecer un argumento.

En 1951, W.V. Quine publicó un artículo con el título «Dos dogmas del empirismo» [4]. La confianza filosófica en la noción de verdad «*a priori*» ha sufrido una erosión ininterrumpida desde ese momento. Quine señaló que muchas cosas que creíamos saber *a priori* han tenido que ser revisadas. Por ejemplo, supongamos que alguien le hubiese sugerido a Euclides que pueden haber dos líneas rectas perpendiculares a una tercera, pero que se *cruzan*. Euclides habría replicado que la imposibilidad de que esto suceda es una verdad necesaria. Pero de acuerdo con la teoría física que aceptamos hoy, esto *sí* sucede. Si la luz que pasa cerca del sol se comporta tal y como lo hace, no es porque viaje en líneas curvas, sino porque la luz sigue viajando en líneas rectas y éstas se comportan de ese modo en nuestro mundo no-euclidiano.

Una vez aceptado este punto, algún filósofo estaba obligado a preguntar «¿Qué queda de la noción de *a priori*?», y Quine lo hizo. (Quine también mostró de forma convincente que las *descripciones empiristas típicas* de la noción de *a priori* —por ejemplo, la noción de «verdad por convención»— eran incoherentes, mas no examinaré sus argumentos.)

Creo que Quine se excede en algunos puntos. Su afirmación «Ningún enunciado es inmune a la revisión», sugiere que para cada enunciado existen circunstancias bajo las cuales su rechazo sería *racional*. Pero no cabe duda de que esto es falso: después de todo, ¿bajo qué circunstancias sería racional rechazar el enunciado «No todos los enunciados son verdaderos», es decir, aceptar el enunciado «Todos los enunciados son verdaderos»? [5].

Aunque Quine exagera las acusaciones contra la noción de *a priori*, tiene razón en un punto: nuestras nociones de racionalidad y de revisabilidad racional no se fijan mediante algún manual de reglas inmutable, ni están inscritas en nuestra naturaleza trascendental (como Kant pensaba) por la excelente razón de que la propia idea de una naturaleza trascendental, una naturaleza que poseemos *nouménicamente* —sin tener en cuenta aquellos sistemas en los que podemos con-

---

[4] «Two Dogmas of Empiricism», publicado por primera vez en *The Philosophical Review*, 1951. Reimpreso en *From a Logical Point of View*, Nueva York, 1961. (*Desde un punto de vista lógico*, Ariel, Barcelona, 1962.)

[5] Discuto el ataque quineano a la noción de *a priori* en «Analyticity Beyond Wittgenstein and Quine», *Midwest Studies in Philosophy*, vol. IV, Minnesota, 1979.

cebirnos histórica o biológicamente— carece de sentido. Y puesto que nuestras nociones de racionalidad y revisabilidad racional son producto del conjunto de nuestra muy limitada experiencia y de nuestra muy falible constitución biológica, es de esperar que hasta los principios que consideramos como *«a priori»* o «conceptuales», o como se nos antoje, resulten necesitar de vez en cuando una revisión, a la luz de experiencias inesperadas o de innovaciones teóricas no-anticipadas. Con todo, esta revisión no puede ser ilimitada: de otro modo se disiparía el concepto de lo que podríamos llamar *racionalidad;* pero, por lo general, no está a nuestro alcance establecer límites. Dejando aparte los casos triviales (por ejemplo, «No todo enunciado es verdadero») no podemos estar seguros de que en *ningún* contexto sería racional desechar un enunciado que se considera como una verdad *necesaria* (y de modo legítimo, *en un contexto determinado*). Hemos de admitir que, en general, las consideraciones de simplicidad, de utilidad global y de plausibilidad pueden llevarnos a renunciar a algo que anteriormente tomábamos por un *a priori,* y que hacer esto es *razonable. La filosofía se ha convertido en anti-apriorística.* Pero una vez hemos reconocido que la mayor parte de las verdades que considerábamos como verdades *a priori* tienen un carácter contextual y relativo, *hemos renunciado también al único buen argumento que existía en contra de la identidad mente-cuerpo.* Los teóricos de la identidad estaban obligados a poner de manifiesto este punto, y obraron en consecuencia. Así que la situación cambió.

He estado haciendo uso de la noción de *propiedad,* aunque creo que hemos llegado a confundir al menos dos nociones de «propiedad» [6]. En relación con la noción más antigua solía emplearse el término «predicado» (por ejemplo, en la célebre pregunta «¿Es la existencia un predicado?»). La otra noción es la que usamos hoy al hablar de «propiedades físicas», «magnitudes fundamentales», etc. Cuando un filósofo tiene en mente la primera, considera frecuentemente que el discurso acerca de propiedades es intercambiable con el discurso acerca de *conceptos.* Según este filósofo, las *propiedades* no pueden ser idénticas a menos que constituya una verdad conceptual el que lo sean; en particular, la *propiedad* de experimentar una sensación con cierto carácter cualitativo no puede ser la misma propiedad que la de hallarse en cierto estado cerebral, ya que los correspondientes predicados no son *sinónimos* (en el amplio sentido de «analíticamente equivalentes»), y el principio de individuación para predicados consiste precisamente en que *ser P* es uno y el mismo predicado que *ser Q* si y solo si *«es P»* es sinónimo de *«es Q».*

No obstante, consideremos la situación que se crea cuando un cien-

[6] Véase «On Properties», capítulo 19 de mi libro *Mathematics, Matter and Method, Philosophical Papers,* vol. I, Cambridge, 1975.

tífico afirma que la temperatura *es* energía cinético-molecular media. A primera vista, este enunciado identifica propiedades. Lo que se afirma es que la *propiedad* de tener una temperatura particular es *realmente* (en algún sentido de «realmente») *la misma propiedad* que tener cierta energía molecular, o (de forma más general) que la *magnitud física* temperatura y la magnitud física energía cinético-molecular media son una y la misma propiedad. Si esto último es correcto, entonces, ya que «*X* tiene tal y tal temperatura» y «*X* tiene tal y cual energía cinético-molecular media» no son oraciones *sinónimas,* aun cuando «tal y cual» sea el valor de la energía molecular correspondiente al valor «tal y tal» de la temperatura, habremos de concluir que el físico llama «magnitud física» a alguna cosa distinta de lo que los filósofos han llamado «predicado» o «concepto».

Siendo más explícito: la diferencia estriba en que mientras que para que los predicados *P* y *Q* sean los mismos se requiere la sinonimia de las expresiones «*X* es *P*» y «*X* es *Q*», esta sinonimia no es un requisito para que la propiedad *P* y la propiedad *Q* sean la «misma» propiedad. Las propiedades, al contrario que los predicados, pueden ser «sintéticamente idénticas».

Si existe algo como la *identidad sintética de propiedades,* entonces ¿por qué no puede ocurrir que la propiedad de hallarse en cierto estado cerebral *sea la misma propiedad* que la de experimentar una sensación con cierto carácter cualitativo (esto es algo que se encuentra muy en la línea del pensamiento de Spinoza), aun cuando ello no constituya una verdad conceptual? De hecho, ¿habrá muchos a quienes les parezca falso *a priori*? En resumen, nos encontramos con una ola de antiapriorismo y con un nuevo sistema para la *identidad sintética de propiedades,* y gracias a ambas, los teóricos de la identidad, y en particular, el funcionalista, parecen entrar en acción automáticamente.

Ahora bien, quiero examinar lo que sucede cuando a estas dos cosas añadimos una tercera. ¿Qué sucede si un filósofo es (1) naturalista antiapriorista, (2) permite que haya algo como una identidad sintética de propiedades y (3) mantiene una concepción realista con respecto a la verdad, en la *línea más dura?* Deseo afirmar que este mismo filósofo se verá enfrentado a serias dificultades epistemológicas.

CEREBROS ESCINDIDOS

Examinemos un determinado tipo de experimento que los neurólogos han venido realizando en los últimos veinte años. Es el célebre experimento de los «cerebros escindidos» o experimento de disociación cerebral.

Quiero traer a colación la relevancia de este tipo de experimento

con respecto a la teoría de la identidad y con respecto a lo que hasta este punto se ha dado por sentado en toda la discusión: la idea de que hay una *correlación*.

Según el modelo del cerebro como un sistema cognitivo semejante a una computadora, el cerebro posee un lenguaje interno (que podría ser innato, o una mezcla de «lenguaje» o sistema de representación innato y lenguaje público). Algunos filósofos han llegado incluso a inventar un nombre para este hipotético lenguaje cerebral, el «mentalés». Consideremos lo que sucede cuando uno experimenta una sensación visual, según este modelo (me tocará en parte inventar mi neurología, ya que no tengo los suficientes conocimientos en este campo, aunque no creo que nadie los tenga realmente). Las cosas podrían suceder así:

Cuando uno experimenta una sensación, se lleva a cabo un «juicio»: el cerebro tiene que «imprimir» algo así como «registrado el color rojo a las 12 en punto». De modo que la cualidad (llamémosle «*Q*») corresponde, entre otras cosas, a *una grabación en mentalés.* Del mismo modo, se da un *input* hacia el centro de procesamiento verbal, conectado con las cuerdas vocales; estas últimas dan cuenta de la capacidad del cerebro para informar, en *lenguaje público,* «experimento rojo en este momento». Puede que el juicio en mentalés tenga que ser transmitido desde una posición a otra antes de que se dé un *input* hacia el centro de habla. También ocurren eventos en el córtex visual (estudiados por los neurólogos Hubel y Wiesel), que imagino destinados a «grabarse en mentalés», así como un proceso verbal. Estos «registros», «*inputs*» y demás acontecimientos pueden tener lugar en diferentes lóbulos del cerebro: si se secciona el *cuerpo calloso,* el lóbulo derecho de la persona (que no posee habla) puede ver (algo) rojo (o al menos dará una señal afirmativa a una pregunta escrita visible sólo para este lóbulo) pero si se le pregunta al sujeto «¿De qué color es la tarjeta?» responderá «No puedo ver la tarjeta»*. Y, finalmente, en algún punto se forman la huella o huellas de memoria (disociable en *memoria a corto plazo y memoria a largo plazo*). Es casi seguro que no existe una cadena causal; probablemente existen ramificaciones y reincorporaciones: es decir, una red causal.

El problema consiste en que la psicología divide los eventos mentales de una forma excesivamente *discreta.* He aquí la sensación de azul: en este momento empieza, en este otro acaba. Pero las redes causales no son discretas. No hay un único evento físico que sea *el* correlato de la sensación.

Si la teoría de la identidad está en lo cierto, el estado sensorial *Q*

* Para una descripción más detallada del experimento de Sperry, véase K. POPPER y J. ECCLES: *El Yo y su Cerebro,* trad. cast. de Carlos Solís, Labor, Barcelona, pp. 359-362 *(N. del T.).*

es idéntico a algún estado cerebral. Un realista metafísico no puede considerar esta identidad como un asunto de convención o de decisión o como si se tuviese un componente convencional idéntico al estado cerebral *Q*. La opinión es que, como *cuestión de hecho*, vivimos en un mundo en el cual lo que experimentamos como caracteres cualitativos de las sensaciones son en realidad *las mismas propiedades* que algunas de las propiedades que nos encontramos en otros campos como propiedades físicas de eventos cerebrales. (O, mejor dicho, en los cuales la propiedad de experimentar una sensación con cierto carácter cualitativo es exacta y realmente la propiedad de hallarse en cierto estado cerebral.)

Detengámonos por un momento y veamos lo que realmente afirma tal punto de vista. Supongamos que estamos fijándonos en la cualidad subjetiva de *rojo* (producida, por ejemplo, al mirar fijamente un disco verde, retirándolo para obtener una post-imagen). Supongamos también que cuando experimento esta cualidad de *rojo*, el estado sensorial en el que me hallo es idéntico a una disyunción de estados cerebrales. El estado sensorial no puede ser idéntico a un estado cerebral máximamente especificado, puesto que sabemos que seguiríamos teniendo la misma experiencia aun cuando anulásemos una neurona, o cualquier otro elemento. Pero la propiedad puede ser disyuntiva, por ejemplo (aunque es bastante implausible): *o bien se están descargando las neuronas pares del área tal-y-tal o bien se están descargando aquellas cuyo ordinal es un número primo*. Habría una descomunal colección de estados neurológicos tales que su disyunción constituiría la propiedad de experimentar *rojo*.

Vayamos ahora un poco más lejos: si se están descargando las neuronas pares del area tal-y-tal, experimento *rojo*. Pero si el cerebroscopio dice «No, están descargándose las neuronas cuyo ordinal es un número primo del área tal-y-tal», también experimento *rojo*. Es decir, no puedo decir en cuál de estos estados cerebrales me hallo. Pero si experimento *rojo*, he de hallarme en uno de ellos. Mas no puedo *distinguirlos*. *Están descargándose las neuronas pares del área tal-y-tal* no es una propiedad observable. Aun sabiendo que la teoría de la identidad es verdadera, no puedo decir, a partir de mis sensaciones, que tengo esta propiedad. Llamemos «P_1» a esta propiedad y «P_2» a la propiedad de que *las neuronas impares del área tal-y-tal se estén descargando*. El estado sensorial es idéntico a la disyunción (P_1 o P_2), siendo ésta, por supuesto, una tercera propiedad. P_1 no es un estado sensorial y P_2 tampoco lo es; sólo su *disyunción* constituye un estado sensorial. En otras palabras, según esta ontología, la *disyunción* de dos propiedades que en sí mismas son *inobservables* puede ser *observable*. Lo que experimento como algo dado de forma simple es sin embargo una complicada función lógica de propiedades inobservables. Esa es la posición.

Es posible que le haya dado una apariencia estúpida. Cierto amigo mío me ha comentado «Supongamos que el único mecanismo que tenemos para detectar muones no distingue entre muones y antimuones. Entonces *muón* no es una propiedad observable, como tampoco lo es *antimuón,* aunque sí lo es su disyunción. Mas ello sólo puede parecer paradójico a quienes conciben la observacionalidad como una noción menos pragmática de lo que en realidad es». No obstante, mi propósito no es ridiculizar esta posición —en realidad constituye un programa de investigación neurofisiológica muy importante y completamente legítimo—, sino dejar claro a qué nos compromete. Voy a argumentar que no es la teoría de la identidad la que de por sí nos conduce a dificultades, sino la teoría de la identidad considerada en conjunción con el realismo metafísico —es decir, considerada en conjunción con lo que he denominado perspectiva «externalista» con respecto a la naturaleza de la verdad.

Se puede eludir el compromiso con esa perspectiva. Por ejemplo, Carnap habría dicho (al menos en cierto período) que el discurso acerca de objetos físicos es en realidad discurso acerca de sensaciones, si bien muy derivado, y que la decisión de afirmar que un particular estado cerebral es idéntico a un estado sensorial *Q* es, en realidad, la decisión de modificar de cierta forma el lenguaje del discurso acerca de propiedades físicas, de cambiar nuestro concepto de la propiedad física en cuestión.

Puesto que el discurso acerca de objetos y propiedades físicas es tan sólo una derivación del discurso acerca de sensaciones, podemos modificar las reglas. Pero este punto de vista no es el del realismo metafísico, al menos en lo que respecta a objetos materiales y propiedades físicas. Quien desee ser un realista metafísico en lo que atañe a *sensaciones,* sin serlo con respecto a los *objetos materiales,* puede adoptar la teoría de la identidad (ya que considera al discurso acerca de objetos materiales como algo flexible) afirmando simplemente «La adopto como un tipo de *convención,* como una estipulación adicional de significado». Ya que los significados no estaban fijados de antemano y ya que existía cierta textura abierta, desaparece el problema consistente en «¿Cómo podemos saber que el estado sensorial es idéntico a esta propiedad y no a alguna otra?». Si lo que *es esta* propiedad es bastante vago, estamos autorizados a postular la identidad simplemente como una especificación de significado. Pero mi debate tiene como interlocutor a alguien que realmente piense que existe un mundo material ahí fuera, y que éste no es una mera derivación del discurso acerca de sensaciones; alguien que realmente crea que hay propiedades físicas, y que sostenga que expresiones tales como «Se están descargando las neuronas en tal-y-tal canal» son predicados que definen nuestras propiedades físicas, y que cualquiera de estas propiedades *o bien es idéntica a este estado sensorial* o bien no lo es.

De modo similar, creo que un filósofo como Daniel Dennett, quien piensa que el discurso acerca de sensaciones es sumamente vago, y no cree que haya una propiedad subjetiva bien definida que consista en hallarse en tal estado sensorial o en experimentar una sensación con tal carácter cualitativo, podría adoptar una teoría de la identidad en la forma de estipulación de significado, aunque lo que se determinaría en esta ocasión sería el significado de los términos psicológicos, y no el de los términos de objetos físicos. Pero, una vez más, la actitud de un realista metafísico radical no sería ésta.

Estoy pensando en un realista metafísico radical que discurra de esta guisa: «Sí, sé en que consiste esta propiedad psicológica (el estado sensorial). La he experimentado, puedo reconocerla. Creo que es la propiedad psicológica a la que me refiero. Sé lo que son P_1 y P_2, y, por ende, lo que es (P_1 o P_2), y el estado sensorial o bien es idéntico a esta propiedad o bien no lo es», tal y como un físico podría decir: «No hay ningún elemento convencional (creo que estaría en un error, dicho sea de paso) en la decisión en favor de que *la temperatura es energía cinético-molecular media,* así que, o bien la temperatura es energía cinético-molecular media o bien es alguna otra propiedad». Este es el punto de vista que quiero examinar.

El problema radica en que, si uno acepta el punto de vista del realista metafísico, existen muchas más posibilidades de las que la gente acostumbra a considerar. La primera posibilidad que se nos ocurre consiste en que el estado sensorial sea idéntico a la propiedad de que ocurran los eventos adecuados en el córtex visual *y* que se tenga adecuadamente grabado el «registro» en «mentalés», *y* que se dé un *input* hacia el centro de habla *y* que se hallen formadas las huellas de la memoria —es decir, el estado sensorial se concibe idéntico a la *conjunción* de todas las propiedades. Pero tan pronto consideramos la posibilidad de la disociación cerebral dejamos de estar seguros de que sea necesaria toda la conjunción. Quizá la sensación sea precisamente cierto evento en el córtex visual (es decir, la propiedad de experimentar una sensación es «realmente» la propiedad de que en el córtex visual tenga lugar cierto evento).

Supongamos que es así, de momento. Ahora bien, supongamos que podemos bloquear el proceso que produce el registro en mentalés, o, al menos, que podemos bloquear el *input* que se dirige hacia el centro de habla. Imaginemos que le hemos mostrado al sujeto una tarjeta roja en el lado izquierdo de su campo visual (de forma que la tarjeta sólo es «visible para el lóbulo derecho», como dicen los neurólogos). El evento en cuestión, que tendrá lugar en el córtex visual, se dará entonces en el lóbulo derecho; pero si le preguntamos al sujeto «¿Ha visto usted algo rojo?», contestará «No».

Ahora bien, por mor de un criterio que empleamos a la hora de decidir si alguien experimenta o no una sensación, el criterio de since-

ridad en los informes verbales, *no nos queda más remedio que admitir que el sujeto no ha experimentado la sensación de rojo*. Y, por lo tanto, tendremos que admitir la «refutación» de la teoría de que *Q* (el carácter cualitativo relevante) es idéntico a esa propiedad del córtex visual. Pero alguien podría objetar «No, de ningún modo ha *refutado* la teoría. Pues, ¿qué tipo de observador nos ofrece? El sujeto tiene el cerebro partido por la mitad». En tanto dispongamos de un observador en *condiciones normales,* la propiedad *Q* es idéntica a esta propiedad del córtex visual. Y los observadores que no están en *condiciones normales* no cuentan. No *pueden* contar.

La dificultad consiste en que existen teorías de la identidad *observacionalmente indistinguibles* [7], y con ello quiero decir que son teorías que conducen a las mismas predicciones con respecto a la experiencia de todos los *observadores que se hallen en condiciones normales.*

Consideremos la teoría que afirma que es imposible tener la sensación de rojo salvo que se dé el *input* hacia el centro de habla. ¿Cómo puede ser probada o refutada? Podríamos pensar que hay un modo, siempre que escindamos el cuerpo calloso y que parte de la memoria no atraviese la unidad de proceso verbal; a saber, primero le preguntamos al sujeto si experimenta la sensación de rojo. Este contesta: «No». A continuación volvemos a coser el cuerpo calloso (¡Un hábil truco, si pudiésemos realizarlo!), y le preguntamos «¿Ha experimentado usted la sensación de rojo?». Podría contestar «Sí, pero esto es un disparate: usted sabe que la he experimentado y aun así me lo pregunta, y yo me he oído a mí mismo contestándole con toda sinceridad que no». (En realidad es más usual que los pacientes «reconcilien» o racionalicen situaciones de este tipo a que las describan como acabo de imaginar.) ¿Demostraría tal informe que *hubo* una sensación de rojo sin que se diese un *input* hacia el centro de habla?

No lo haría. Si Daniel Dennett (quien alguna vez mantuvo la opinión de que la sensación *es* el *input* hacia el centro de habla, o una

[7] La noción de «indistinguibilidad observacional» fue introducida en varios artículos sobre la teoría del espacio-tiempo por CLARK GLYMOUR y DAVID MALAMENT en *Foundations of Space-Time Theories,* Earman, Glymour and Stachel (eds.), *Minnesota Studies in the Philosophy of Science,* vol. VIII, Minnesota University, 1977. Un problema análogo en esta teoría es la existencia de «posibles» espacio-tiempos (es decir, espacio-tiempos autorizados por la teoría de la relatividad) que difieren en sus propiedades topológicas globales, pero en los cuales los observadores tendrían exactamente las mismas experiencias. Tales ejemplos se despachan a menudo sobre la base de que las «consideraciones de simplicidad» nos dirían en qué espacio-tiempo estamos viviendo; el problema (como señala Malament) es que la teoría física (la relatividad general) *no dice* que vivamos en el espacio-tiempo *más simple* de los compatibles con sus leyes.

opinión muy cercana a ésta[8]) deseara reconciliar el informe de este sujeto con su teoría, todo lo que tendría que decir es: «No niego que en la última ocasión tuviese lugar el evento psicológico consistente en *recordar que se ha experimentado antes la sensación.* Lo que niego es que la *sensación* tuviese lugar en la ocasión anterior». En una u otra teoría, el sujeto tiene *más tarde* la experiencia de recordar *correcta o erróneamente* que experimentó antes la sensación de rojo.

En este punto el desacuerdo es real. Muchos neurólogos creen que el lóbulo derecho de los pacientes con el «cerebro escindido» es «consciente». En efecto, esto equivale a afirmar que a veces hay una sensación de rojo, a pesar de que no se dé ningún *input* hacia el centro de habla. La expresión que frecuentemente se utiliza es «Existen dos *loci* de consciencia». Al menos un famoso neurólogo, Eccles, mantiene que, pese a todo, el lóbulo derecho disociado (o el izquierdo, en el caso de los pacientes que tienen el centro de habla en el derecho) *no* es consciente. Según Eccles, la consciencia es *unitaria;* que el lóbulo disociado pueda «simular» una conducta consciente no demuestra que sea un segundo «*locus* de consciencia».

Tampoco nos ayudará apelar a máximas metodológicas, por ejemplo, «Elija usted la teoría más simple»: pues no parece que exista ningún tipo relevante de «simplicidad» que sea patrimonio de la teoría «unitaria» y del cual carezca la teoría de los dos «*loci*» (y vicecersa). La teoría de los dos «*loci*» es más simple en un aspecto: afirma que ciertas capacidades conductuales (que posee el lóbulo derecho, aun cuando no posea habla) son suficientes para la consciencia, y esto casa con el hecho de que calificamos como conscientes a los animales (los cuales tampoco poseen habla). Pero existen muchas *desemejanzas* entre un animal con el cerebro intacto, cuyos procesos cerebrales aún están *integrados* (aunque no involucren el habla), y una parte de un «cerebro escindido». Si el caso no nos tocase tan de cerca, si no tuviésemos una tendencia tan acusada hacia el realismo metafísico con respecto a las *sensaciones*, ¿no estaría más de acuerdo con nuestras intuiciones metodológicas considerarlo como un caso a *legislar,* en vez de una cuestión sobre la que disputar?

En suma, existen varias teorías de la identidad observacionalmente indistinguibles. Si el teórico de la identidad está en lo cierto, da la impresión de que no hay forma humana de saber de que modo está en lo cierto, de saber cuál es el estado cerebral que es idéntico al estado sensorial (o está *correlacionado con* éste).

Thomas Nagel[9] ha defendido la plausible afirmación de que es

[8] DENNETT presenta su modelo de conciencia en «Towards a Cognitive Theory of Consciousness», reimpreso en su libro *Brainstorms,* Bradford Books, 1978.

[9] «What is it like to be a bat?», reimpreso en N. Block, *op. cit.*

imposible imaginar cómo siente un murciélago. Pero, *¿por qué motivo* ha de ser una afirmación plausible? Hace algunos años leí un delicioso libro de Donald Griffin sobre los murciélagos. Llegué a darme cuenta de que los murciélagos no son básicamente diferentes de los demás mamíferos. En general pensamos que está en nuestra mano imaginar qué sensaciones tienen nuestros perros y nuestros gatos. ¿Cuál es la dificultad con respecto a los murciélagos?

Bien, los murciélagos pueden oír sonidos que son más agudos en varias octavas que los que nosotros podemos oír. No puedo imaginar cómo siente un murciélago en el sentido de que no puedo imaginar en qué consiste la sensación de localización mediante el eco. Pero ¿es preciso que sea *tan* difícil? Yo mismo solía ser capaz de oír sonidos una octava más agudos que los sonidos más altos que puedo oír ahora, en mi madurez. Pero el tono subjetivo no ha cambiado: los sonidos más agudos que puedo oír ahora pueden ser una octava más graves en cuanto a tono *objetivo* que los sonidos más agudos que podía oír cuando tenía diez años y, sin embargo, poseen la misma propiedad tenue y chirriante que siempre tuvieron para mí los sonidos que se hallan en el umbral extraauditivo por ser demasiado agudos. Quizá sea así como un murciélago oye un sonido que es cinco octavas más alto que aquellos que nosotros podemos oír: como un chirrido corto y agudo.

Ahora bien, imaginemos un debate entre dos filósofos, o entre dos psicólogos, uno de los cuales afirma que *ninguna cualidad* experimentada por un murciélago se parece en algo a las cualidades experimentadas por los seres humanos. Los *qualia* del murciélago son inimaginablemente distintos de los *qualia* humanos. Nunca seremos capaces de imaginar en qué consiste sentir como lo hace un murciélago (ni siquiera como un perro o un gato). Podemos imaginar al otro filósofo replicando : «Lo que usted afirma no tiene sentido. Quizá no pueda imaginar *algunas* sensaciones del murciélago; pero también es probable que no pueda imaginar algunas sensaciones de otros seres humanos (por ejemplo, algunas sensaciones del sexo opuesto), y, sin embargo, esto no significa que conciba el espacio psicológico de esos seres humanos como si fuera inimaginablemente distinto del mío. ¿Por qué no debo pensar que el campo visual del murciélago, por ejemplo, es muy parecido al mío? (N.B. Al contrario de lo que se piensa, los murciélagos ven perfectamente.) Permitiendo algunos ajustes en la óptica del ojo del murciélago o en la acústica de su oído, de forma que caiga dentro de un rango que coincida con el mío, su oído es como el mío y sus dolores como los míos». Ahora bien, ¿cómo podríamos *resolver* esta disputa?

Ya que tanto el número de neuronas como su disposición son distintos (el centro acústico del cerebro del murciélago se amplía hasta convertirse en *7/8* partes del cerebro), las propiedades a nivel neuro-

lógico, especificadas al máximo —cuántas neuronas se descargan y dónde lo hacen— que son idénticas a una *cualidad* en el caso del murciélago (en el supuesto de que la teoría de la identidad sea correcta), no pueden ser literalmente las mismas que las propiedades que son idénticas a alguna *cualidad* en el caso de un ser humano. ¿O sí pueden serlo? Supongamos que cuando un murciélago experimenta cierta sensación visual (producida al ver objetos rojos), el cerebro del murciélago tiene la propiedad disyuntiva (P_1 o P_2), donde P_1 y P_2 son estados de su cerebro, máximamente especificados. (De hecho sería una propiedad disyuntiva mucho más complicada, con miles de casos, pero permítasenos simplificar.) Supongamos también que cuando experimento cierta sensación visual (producida al ver objetos rojos), mi cerebro tiene la propiedad disyuntiva (P_1' o P_2'). Consideremos las dos teorías siguientes: (1) el carácter cualitativo de la sensación del murciélago (llamémosle «*rojo$_m$*») es idéntica a *(o al menos está en correlación con)* la propiedad disyuntiva (P_1 o P_2) y el carácter cualitativo de la sensación humana (llamémosle «*rojo$_h$*») es idéntico a (*o al menos está en correlación con*) la diferente propiedad (P_1' o P_2'). (2) El carácter cualitativo de la sensación del murciélago es *idéntico* al carácter cualitativo de *mi* sensación (es decir, *rojo$_m$* = *rojo$_h$*) y ambos son *idénticos* a (o están correlacionados con) la propiedad más compleja (P_1 o P_2 o P_1' o P_2').

Según la primera teoría, el murciélago y yo tenemos experiencias distintas, mientras que según la última tenemos la misma experiencia; sin embargo, estas dos teorías conducen a las mismas predicciones con respecto a lo que experimentarán los observadores *humanos,* normales y anormales. Una vez más, son dos teorías observacionalmente indistinguibles.

¿Nos ayudarán las máximas metodológicas («elija usted la teoría más simple»)? Sigue sin estar nada claro que puedan hacerlo. Ned Block ha señalado que mientras la primera teoría es más simple en un aspecto (la *cualidad* se identifica con una propiedad más simple en *cada caso*), la segunda lo es en otro (es «no-chovinista»: consiente que se puedan experimentar las cualidades que nosotros experimentamos sin que se tenga necesariamente nuestra constitución física).

Una vez más, carecemos de principios que determinen cuál de las dos teorías es preferible. En realidad, ¿hay alguna razón para creer que existan o deban existir tales principios? ¿Por qué no abandonar, como Wittgenstein nos recomendó, nuestro realismo metafísico con respecto a las sensaciones y con respecto al predicado «es la misma que» (en tanto que aplicado a las sensaciones) y tratar este caso también como un caso a *legislar,* y no como un caso sobre el que disputar?

Por último, quiero exponer tres teorías que con toda seguridad son falsas, pero cuyo rechazo es difícil o imposible en el caso de que el

realismo metafísico sea correcto. Son éstas : (1) *rojo*$_h$ es, después de todo, un estado funcional o casi-funcional, a saber, el estado consistente en hallarse en cualquier estado material (por ejemplo, físico) que desempeñase en nuestras vidas con anterioridad el rol funcional de señalar la presencia de rojo objetivo. (2) Las rocas tienen *qualia* (es decir, en las rocas tienen lugar eventos cualitativamente similares a las *sensaciones visuales*). (3) Las naciones son conscientes.

Consideremos primero (1). Recordemos el argumento que utilicé para mostrar que *rojo*$_h$ no podía ser un estado funcional. El argumento consistía en que si identificamos *rojo*$_h$ con el estado funcional de *hallarse en cualquier estado material (por ejemplo, un estado cerebral) que normalmente señale la presencia del rojo objetivo,* entonces yo no podría haber sufrido una inversión del espectro (al menos en el caso de la «amnesia»), ya que me hallo en ese estado funcional cuando veo algo objetivamente rojo tanto *antes* como *después* de la inversión del espectro (dejando tiempo para que tenga lugar un ajuste lingüístico, y, si fuese necesario, postulando un ataque de amnesia). Pero, según la posición del realismo metafísico, es sin duda posible que haya sufrido una inversión del espectro (aun cuando no lo recuerde debido al ataque de amnesia). El caso es más dramático todavía si no he tenido un ataque de amnesia y sí *recuerdo* que mi espectro se ha invertido; pero aun así, si los ajustes lingüísticos han sido automáticos, hay un sentido en el que lo que solía ser «la sensación de verde» desempeña ahora el rol funcional de «señalar la presencia de *rojo* objetivo en el entorno».

Este argumento sólo prueba que *rojo*$_h$ no es idéntico al estado funcional señalado. No prueba que no sea idéntico a un estado funcional más complejo, tal como *hallarse en cualquier estado material que con anterioridad materializase en nuestras vidas el susodicho estado funcional.* Se podría objetar que esta propiedad es bastante extraña: una complicada función lógica de propiedades funcionales. Pero ¿por qué razón es menos verosímil que una complicada función lógica de propiedades funcionales sea idéntica a *rojo*$_h$ a que lo sea una disyunción de complicadas propiedades físicas? ¿Acaso el mundo prefiere disyunciones de propiedades físicas a conjunciones de propiedades funcionales?

Consideremos (2). Sea P_3 la propiedad de ser una roca, y consideremos la hipótesis de que *rojo*$_h$ es idéntico a la propiedad disyuntiva (P_1 o P_2 o P_1' P_2' o P_3). Las rocas poseen esta propiedad siempre, por supuesto. Así que, según esta hipótesis, en las rocas tienen lugar continuamente eventos con el carácter cualitativo *rojo*$_h$. (No están experimentando rojo en el sentido *funcional,* pero en ellas se da continuamente un evento con el mismo carácter cualitativo que tiene el evento que desempeña en nosotros el rol funcional de ser la sensación de rojo.) O consideremos la hipótesis según la cual las rocas tie-

nen diferentes *qualia* en diferentes ocasiones. O, al menos, consideremos la hipótesis de que alguna de estas hipótesis es correcta (sin especificar cuál). Podríamos decir, «Bueno, pero estas hipótesis son disparates». En efecto, lo son. Pero todas y cada una de ellas conducen a las mismas predicciones, para todos los seres humanos, a las que conducen la teoría «sana». Ninguna de ellas puede ser excluida sobre una base observacional o experimental, puesto que todas ellas son observacionalmente indistinguibles, desde la concepción más estándar.

Podríamos pensar que estas teorías pueden excluirse apelando al principio metodológico: *no se ha de atribuir una propiedad a un objeto sin ninguna razón.* Este principio no dice, por supuesto, que tales teorías son *falsas* (a veces no tenemos ninguna razón para creer cosas que resultan ser verdaderas), pero al menos afirma que está *justificado el considerarlas falsas.* Pero ¿seguro que no existe ninguna razón para sostener la menos específica de estas teorías (la teoría que afirma que alguna de estas teorías es correcta y las rocas tienen *qualia*)? ¿Qué ocurre con el argumento de que si *nosotros* tenemos *qualia* y el fisicalismo es verdadero (y muchos filósofos creen que hay multitud de buenas razones para aceptar el fisicalismo), entonces hay al menos un objeto físico en el cual tienen lugar eventos con carácter cualitativo: así que por qué no han de tener lugar tales eventos en *todos* los objetos físicos? Esta posibilidad se cerraría si pudiésemos mostrar que existe algo en la propia *cualidad* que le *exige* tener el «rol» funcional particular que tiene para los seres humanos; pero los creyentes en los *qualia* como objetos metafísicamente reales nos dicen que es esto precisamente lo que no podemos hacer.

Consideremos (3), por último (aunque no por ello es la menos importante). Consideremos la hipótesis de que el dolor es idéntico a un estado funcional adecuado, que puede manifestarse también en los organismos institucionales o en las naciones. En otras palabras, supongamos que cuando en España se declara que «España está afligida por...», realmente *lo está.* Nunca lo sabríamos, claro. Quizá el lector encuentre en este momento interesante y algo divertido el que un grupo pueda comportarse de manera semejante a como se comporta, cuando siente dolor, algo que realmente siente dolor; pero el lector no cree que España realmente sienta dolor. Según la hipótesis, el lector estaría equivocado: el espíritu nacional realmente estaría sufriendo dolor.

Esta hipótesis guarda relación con una interesante polémica en filosofía de la mente. A los funcionalistas (entre los que me incluyo) les gusta emplear el siguiente argumento «antichovinista»: las diferencias entre un robot y un ser humano, en cuanto a organización funcional, pueden reducirse, en principio, a pequeños detalles físico-químicos. Hasta podríamos encontrar un robot cuyo nivel funcional fuese correspondiente al nuestro. (E incluso podría tener un cuerpo

«de carne y hueso», además de cerebro.) Ahora bien, salvo que seamos unos «chovinistas del hidrógeno-carbono» y pensemos que el hidrógeno y el carbono son intrínsecamente más conscientes, ¿por qué no decir que este robot es una persona cuyo cerebro resulta tener más metal y menos hidrógeno y carbono?

Este argumento ha provocado la siguiente réplica: «Bien, supongamos que en lugar de estos chismes electrónicos (neuronas electrónicas ensambladas en los mismos circuitos en los que están las neuronas humanas) lo que hay es gente minúscula, pequeños boy-scouts». Ni siquiera tenemos que imaginar que estos seres minúsculos conocen la función de todo el esquema, ni que ven otra cosa que una habitación tenuemente iluminada, o un montón de habitaciones tenuemente iluminadas, en las que se pasan notas unos a otros. (Su tiempo habría de transcurrir muy rápido en relación al «nuestro», por supuesto.) Podrían ser obreros alienados. «Ahora bien», continúa la réplica, «no diríamos que tal cosa es ''consciente'', ya que sabemos que, en realidad, consta sólo de seres minúsculos moviendo su cuerpo. Y esto muestra que una organización funcional adecuada (como la nuestra) no basta para justificar la aplicación de predicados como ''consciente''».

Una réplica a esta réplica (la que de hecho aduje) es negar que el «robot hidra-cefálico» (así se le ha llamado) tenga la misma organización funcional que nosotros tenemos. Pero podría haber contestado con una réplica más radical. Podría haber preguntado «¿Qué *nos prohíbe* afirmar que el robot hidra-cefálico es consciente? Si el primer argumento es correcto (y creo que lo es), si el robot que posee un cerebro positrónico puede ser consciente, ¿por qué motivo el hecho de que las neuronas del robot hidra-cefálico son *más* conscientes significa que ese ente, en su totalidad, es *menos* consciente? Después de todo, somos, hasta cierto punto, una sociedad de pequeños animales. Nuestras células son, hasta cierto punto, pequeños animales individuales. Y quizá tengan sentimientos, aunque estos sean ínfimos, ¿quién sabe?». Ahora bien, si jugamos esta baza, si decidimos que el robot hidra-cefálico es consciente (pese a que sus neuronas sean boy-scouts), ¿por qué no lo puede ser España?

No pretendo afirmar que España tenga la misma organización funcional que el *homo sapiens.* Está claro que no. Pero existen muchas semejanzas. España tiene órganos de defensa. Tiene órganos de ingestión, se alimenta de petróleo y cobre, etc. Excreta (polución) en vastas cantidades. ¿Acaso no es su organización funcional tan semejante a la de un mamífero como lo es la de una molesta mosca, a la que *sí* atribuimos dolor?

¿EN QUE MEDIDA ESTA BIEN-DEFINIDO EL «CARACTER CUALITATIVO»?

Por el momento no hemos puesto en duda lo que significa tener dos experiencias que comparten el mismo «carácter cualitativo». Sin embargo, esto no ocurre ni siquiera a nivel introspectivo. En primer lugar, lo que nos parecen nuestras experiencias depende de forma notoria de nuestras conceptualizaciones previas, como muestra el que digamos que vemos una mesa circular aun cuando la observemos desde un ángulo.

Este último caso plantea la cuestión, discutida por psicólogos y filósofos desde el siglo diecinueve, de si tenemos «sense data elípticos» y *pensamos* que son circulares (salvo que estemos suficientemente entrenados en la práctica de la introspección) o tenemos «Gestalts» circulares y pensamos que son elípticos debido a nuestro conocimiento de la teoría óptica. Podemos tener experiencias que se ajusten a cada una de las descripciones; por si fuera poco, muchas experiencias se ajustarán a ambas. Y tampoco es probable que la neurología resuelva esta disputa: es indudable que la imagen elíptica de la retina da lugar a eventos en el cerebro, pero si identificamos *éstos* con «la sensación visual» el resultado puede ser algo parecido al típico relato sobre «sense data elípticos *plus* inferencias inconscientes»; el carácter judicativo de la experiencia («veo la superficie circular de una mesa») corresponde igualmente a «registros» e «*inputs*» en el cerebro, pero si identificamos éstos con la sensación visual, el resultado puede ser otro relato según el cual no tenemos sense-data elípticos salvo que *juzguemos* que algo parece elíptico. ¿Por qué no afirmar que ambas versiones son legítimas? Como Goodman afirma con respecto al caso de un sujeto al que se le pide que describa el movimiento aparente

> Lo que podemos hacer es especificar el tipo de términos, el vocabulario, que debe utilizar, pidiéndole que describa lo que ve, haciendo uso de términos perceptivos o fenoménicos más bien que de términos físicos. Depare o no diferentes respuestas, este procedimiento arroja una luz completamente distinta sobre lo que sucede. La necesidad de especificar el instrumento que debe utilizarse para configurar los hechos convierte en una insensatez cualquier tipo de identificación de lo físico con lo real y de lo perceptivo con la mera apariencia. Lo perceptivo es una versión bastante distorsionada de los hechos físicos en la misma medida en que lo físico es una versión sumamente artificial de los hechos perceptivos [10].

Si veo un mantel rojo en dos ocasiones diferentes a lo largo del día, ¿tengo la *misma* sensación de rojo? ¿O tengo sensaciones distintas y no me doy cuenta de la diferencia (salvo que sea un pintor)?

---

[10] *Ways of Worldmaking*, pp. 92-93.

El caso de la *acomodación* es especialmente difícil de resolver. Si un sujeto suele llevar lentes que invierten la imagen, después de un tiempo las cosas le volverán a parecer normales. ¿Se han restablecido los «*sense-data*» solamente con chasquear los dedos? ¿O es que el sujeto se ha acostumbrado a los *sense-data* alterados y ha reinterpretado «arriba» y «abajo»? Es muy probable que el sujeto sea incapaz de decir hasta qué punto se ha restablecido la normalidad, o cuál de estas cosas ha ocurrido. (Los lectores que lleven lentes bifocales, como yo las llevo, pueden hacerse la pregunta: ¿no parece distinta la mitad inferior del campo visual, pese a que uno no se dé cuenta de la diferencia?) Mientras existan transformaciones a las que los sujetos no se acomodan *nunca* (de hecho, sólo lo conseguimos en relación con cambios simples), y sostengo que *no* nos acomodaríamos a una inversión del color, el fenómeno de acomodación arroja dudas sobre el grado en que «el mismo caracter cualitativo» constituye una noción bien-definida.

REALISMO CON RESPECTO A LOS *QUALIA*

Hemos examinado un conjunto de dificultades escépticas que no pretenden mostrar que la teoría de la identidad o la teoría de la correlación son erróneas (advirtamos que se pueden señalar tantas dificultades para la teoría de la «correlación» como para la teoría de la identidad), sino que, si son verdaderas, existe un gran número de sistemas alternativos para especificar todos los pormenores, de modo que nunca podemos saber cuál de ellos es el verdadero. E ignorar esto significa ignorar la respuesta a muchos problemas tradicionales planteados por los escépticos, tales como si las rocas y otros objetos inanimados tienen *qualia,* si los murciélagos y otras especies tienen o no el mismo tipo de *qualia* que tenemos nosotros, si los grupos pueden sentir dolor, etc.

Pero ¿qué razones puede tener un filósofo para pensar que es una posibilidad lógica el que una roca sienta dolor (es decir que «en» la roca tenga lugar un evento con el mismo «carácter cualitativo» que tiene el dolor humano)? Quizá Russell nos ofrezca alguna clave para desentrañar la naturaleza de este tipo de realismo metafísico. Russell era realista con respecto a los *qualia* y con respecto a los universales. Además, consideraba que los *qualia* eran los universales paradigmáticos. Un universal es, sobre todo, una forma en la que las cosas pueden ser semejantes; y opinaban que las semejanzas cualitativas entre las sensaciones de un mismo sujeto eran los ejemplos epistemológicamente más sencillos y fundamentales de «modos de semejanza entre las cosas». Los *qualia,* para Russell, son los *universales par excellence.*

Un realista tradicional, sin embargo, concibe un universal como

algo completamente bien-definido: las *palabras* pueden ser vagas, pero los universales no. (Una palabra vaga es vaga porque representa a un conjunto vago de conceptos, según dijo Gödel en cierta conversación; pero los conceptos están perfectamente definidos.)

De modo que, si los *qualia* son universales y los universales son por su naturaleza algo bien-definido, debe estar perfectamente-definido si cualquier cosa o evento —incluyendo la mitad de un cerebro escindido, una roca, una nación o un grupo, o cualesquiera eventos que tengan lugar en estos últimos— manifiesta o no un *quale* determinado. Y si el *quale* se concibe como algo totalmente independiente del *rol* funcional que desempeña, si se estima que es completamente contingente que el carácter cualitativo de la sensación de rojo sea el carácter cualitativo de algo que desempeña ese particular *rol* funcional, entonces parece ser una posibilidad lógica que el cerebro escindido o la roca tengan ese *quale.*

Un filósofo que comparta mi deseo de negar que todas y cada una de estas posibilidades tengan sentido (aunque pueda tenerlo alguna de ellas —*existe* la tentación de considerar el lóbulo derecho como un «*locus* de consciencia», y he sugerido que sería legítimo hacerlo) ha de dejar bien claro que no se adhiere a ninguna forma de behaviorismo. Decir que los *qualia* no son entidades bien definidas no es lo mismo que decir que no existen, que todo lo que existe es la conducta, o cualquier otra afirmación de esta índole. Muchas nociones son vagas y, con todo, poseen ciertos referentes claros. La noción de *casa,* por ejemplo, es una noción incorrectamente definida en el caso de los iglúes (¿consideramos que un iglú es una casa?), en el caso de los refugios de los indios navajos *(hogans)* y quizá también en otros casos. Del mismo modo, el hecho de que no tengamos medios para ponderar y decidir si el lóbulo derecho es «consciente» no significa que no existan seres conscientes.

La semejanza cualitativa está definida hasta cierto punto: si experimento una sensación de rojo seguida por una sensación de verde, sé que he tenido sensaciones que no son semejantes (y lo sé sin necesidad de comparar sus *roles* funcionales), y si he experimentado una sensación de rojo seguida por la «misma» sensación de rojo, sé (con la vaguedad que comentamos más arriba) que he experimentado sensaciones semejantes. Pero, para alguien que mantenga una perspectiva «internalista» con respecto a la verdad, no se sigue que tenga que cuestionarse en todos los casos si dos sensaciones (ni siquiera dos eventos arbitrarios) son cualitativamente semejantes o no.

Sea E el evento consistente en experimentar una sensación particular en un tiempo particular y E' algún evento físico en una roca. La suposición de que el carácter cualitativo de E (digamos $rojo_h$) podría ser idéntico o correlativo a una propiedad como (P_1 o P_2 o P_3) (donde P_3 es la propiedad de *ser una roca*) ofende a cualquier sensi-

bilidad humana mínimamente sensata. La suposición de que *E* y *E'* pueden ser eventos «cualitativamente semejantes» es absurda. Ya hemos discutido una explicación de este absurdo: la hipótesis es absurda porque viola la máxima metodológica «no adscriba usted propiedades a un objeto sin tener alguna razón». Pues bien, aunque esta explicación fuese eficaz, estaría lejos de excluir la posibilidad de que las rocas tengan *qualia* (o de ofrecernos una razón de por qué es incoherente la idea de que sí los tienen). Pero si esa transgresión es todo lo que tienen de erróneo las «hipótesis» que afirman que las rocas tienen *qualia,* entonces estamos en posición de afirmar: *por lo que sabemos, es posible que las rocas tengan* qualia, *aunque a priori es sumamente improbable.*

En realidad, la incoherencia de la hipótesis que afirma que las rocas tienen *qualia* es análoga a la incoherencia de la «hipótesis» de los cerebros en una cubeta. Como esta última, presupone una teoría mágica de la referencia. Cualquier ser humano mínimamente sensato consideraría que *E* y *E'* son tan poco semejantes que ni siquiera puede plantearse la cuestión de su «semejanza cualitativa» (en el sentido en que dos sensaciones pueden ser semejantes, esto es, afectándonos del mismo modo). Pero el realista metafísico, a pesar de que no niega esto último, piensa que *E* y *E' podrían ser semejantes aun cuando sea un «disparate» pensar en esta posibilidad lógica.* Y lo piensa porque se halla bajo la ilusión de que el hecho de experimentar la sensación de marras, con su carácter cualitativo, con «su modo de afectar», con los pensamientos y juicios que la acompañan, ocasiona de algún modo que la expresión «el modo en que me afecta esta sensación» (o algún otro sustituto técnico, por ejemplo «el carácter cualitativo de esta sensación», o «$rojo_h$», o «*este quale*») se refiera a un «universal» definido, esto es, a una propiedad absolutamente bien definida de acontecimientos metafísicos individuales. Pero las cosas no son así.

Si efectivamente hubiera robots que fueran funcionalmente isomórficos a nosotros, y trabajásemos, discutiésemos e incluso entabláramos amistad con algunos de ellos, no dudaríamos ni por un momento que fuesen conscientes. (Aún así, podríamos quedar perplejos ante la cuestión de si sus *qualia* son como los nuestros; pero esta cuestión no se plantearía más a menudo que la cuestión de si los perros o los murciélagos tienen los *mismos qualia* que nosotros.)

Supongámos que nos topamos con robots hidra-cefálicos. (Imaginemos que en realidad se *han desarrollado* en alguna parte, gracias a algún proceso biológico, tal y como los animales simbiontes se desarrollan sobre la tierra.) ¿Cuáles serían nuestros sentimientos hacia ellos?

Aunque el caso es tan extraño que uno no puede estar seguro de nada, parece que incluso en este caso (si nuestra interacción se da con el robot entero, y sólo raras veces con sus «neuronas» conscientes —los

«boy-scouts» de mi relato—) podríamos comenzar a atribuirles consciencia; pero probablemente, nuestras opiniones estarían divididas. Si llegáramos a estar seguros de que los robots hidra-cefálicos son conscientes, ¿podríamos comenzar a estarlo, aunque siempre con ciertos escrúpulos, de que España lo es? No lo sé.

Con respecto a todos estos casos, recomiendo con énfasis la perspectiva que niega que en este punto exista algo *oculto,* algún hecho nouménico consistente en la mismidad real de las sensaciones o en la consciencia real de las entidades. Sólo existen hechos empíricos obvios: las rocas y las naciones son enormemente diferentes de las personas y de los animales, los distintos tipos de robots pertenecen a la clase de los objetos, etc. Las rocas y las naciones *no son* conscientes, y, con respecto a la noción de consciencia, esto es algo que consideramos como un hecho.

Esta concepción es tan perturbadora porque hace que nuestros estándares de aceptabilidad racional, justificación, y, por último, de verdad, dependan de estándares de semejanza que sin duda son producto de nuestra herencia cultural y biológica (por ejemplo, que hayamos interactuado o no con «robots inteligentes»). No obstante, ocurre algo parecido con respecto a buena parte del lenguaje que utilizamos en nuestra vida cotidiana, por ejemplo, con respecto a palabras como «persona», «casa», «nieve», y «marrón». Un realista que aceptase esta resolución de los enigmas acerca de los *qualia,* probablemente diría que «los *qualia* no existen realmente» o que los *qualia* pertenecen a nuestro «sistema conceptual de segundo orden». Pero ¿de qué sirve una noción de «existencia» que coloca a las *casas* en el lado de lo *no-existente*? Nuestro mundo es un mundo humano, y la respuesta a qué cosas son conscientes o no, o a qué cosas experimentan sensaciones o no, o a qué cosas son cualitativamente semejantes o no, depende en última instancia de nuestros juicios humanos con respecto a la semejanza y a la diferencia.

5. DOS CONCEPCIONES DE LA RACIONALIDAD

En los capítulos precedentes he estado hablando de racionalidad y de «aceptabilidad racional». Sin embargo, la racionalidad no es algo de lo que se pueda dar cuenta fácilmente.

El problema es análogo en otras áreas. Hace algunos años estudié el comportamiento de los términos de los géneros naturales, *oro*, por ejemplo, y llegué a la conclusión de que la extensión del término no se determina simplemente mediante una «batería de reglas semánticas» u otras normas institucionalizadas. Las normas pueden determinar que ciertos objetos son *ejemplos paradigmáticos* de oro, pero no determinan la plena extensión del término. Ni siquiera es imposible que un ejemplo paradigmático de oro resulte no ser realmente oro, y debería serlo si «ser de oro» estuviese *definido* simplemente por normas.

Estamos dispuestos a considerar que algo pertenece a un determinado género, a pesar de que los *test* de que disponemos no basten para demostrarlo, si ese algo resulta tener en todo caso la misma naturaleza esencial que (o, de un modo más vago, si es «suficientemente semejante» a) los ejemplos paradigmáticos (o que la gran mayoría de éstos). La naturaleza esencial, o la semejanza suficiente, depende tanto del género natural como del contexto (el té helado puede ser «agua» en un contexto, pero en otros no); mas lo que cuenta para que algo sea oro es su composición última, puesto que (desde la Antigua Grecia) se ha pensado que determinar el comportamiento legal de una substancia consiste en determinar tal composición. A no ser que afirmemos que el significado que los griegos daban a su palabra *chrysos* abarcaba todo aquello que tuviese la misma naturaleza esencial que los ejemplos paradigmáticos, ni su búsqueda de nuevos métodos para la detección de oro falso (búsqueda que condujo a Arquímedes a la prueba de densidad) ni sus especulaciones físicas tendrían sentido.

Es tentador adoptar la misma línea con respecto a la racionalidad, y afirmar que lo que determina que una creencia sea racional no son las normas de racionalidad de tal o cual cultura, sino una *teoría ideal* de la racionalidad, una teoría que nos ofrecería las condiciones necesarias y suficientes para que una creencia sea racional en las circunstancias relevantes y en cualquier mundo posible. Esta teoría tendría que *dar cuenta* de los ejemplos paradigmáticos de racionalidad, del mismo modo que una teoría ideal del oro lo hace con los ejemplos paradigmáticos de oro, pero podría ir más allá de éstos, proporcio-

nándonos criterios que nos permitirían entender casos que en la actualidad no podemos desentrañar por completo, tal y como nuestra actual teoría del oro nos permite entender casos que ni el más brillante de los antiguos griegos podría haber entendido. La propuesta de considerar «racional», «razonable», «justificado», etc., como términos de géneros naturales, tropieza con la dificultad general de que las perspectivas de *encontrar* generalizaciones poderosas con respecto a todas las creencias racionalmente aceptables no son muy halagüeñas. Existen poderosas leyes obedecidas por todas las instancias de oro, y esto posibilita (cuando las conocemos) describir el oro como la sustancia que obedecerá tales leyes; pero ¿tenemos alguna probabilidad de encontrar poderosas generalizaciones universales obedecidas por toda instancia de «creencia racional justificada»?

Que haya pocas posibilidades no significa que *no* existan analogías entre la investigación científica sobre la naturaleza del oro y la investigación moral y filosófica. En ética, por ejemplo, comenzamos con juicios acerca de la corrección o incorrección de los actos individuales («informes de observación», por así decirlo) y vamos formulando gradualmente máximas (y no generalizaciones sin excepción) basadas en esos juicios, que a menudo van acompañadas por razones o ejemplos ilustrativos, como por ejemplo «Sea usted amable con un extranjero, pues ya sabe lo que significa ser extranjero en Egipto» (una generalización de «bajo nivel»). Estas máximas afectan y alteran a su vez nuestros juicios sobre casos individuales, de modo que pueden aparecer nuevas máximas que suplementen o modifiquen las anteriores. Después de miles de años de dialéctica entre máximas y juicios sobre casos individuales, un filósofo puede llegar más lejos y proponer una concepción moral (una «teoría»), que puede alterar tanto máximas como juicios singulares, y así sucesivamente.

Podemos encontrar el mismo procedimiento en toda la filosofía (disciplina que es casi coextensiva con la teoría de la racionalidad). En un artículo que publiqué hace algunos años [1] describía los *desiderata* que debía satisfacer un sistema moral, siguiendo a Grice y a Baker, e incluía (1) el *desideratum* de que las suposiciones básicas contengan un *llamamiento amplio.* (2) El *desideratum* de que el sistema moral sea capaz de resistir la crítica racional, (3) el *desideratum* de que la moralidad recomendada pueda *llevarse a cabo.*

La manera de perfeccionar nuestra comprensión de la naturaleza

[1] «Literature, Science and Reflection», *New Literary History,* vol. VII, 1975-76, reimpreso en mi libro *Meaning and the Moral Sciences,* Routledge and Kegan Paul, 1978.

de la racionalidad —la única manera que conocemos— consiste en perfeccionar nuestras concepciones filosóficas de la racionalidad. (Un proceso interminable, pero así es como ha de ser.) Es bastante sorprendente que los *desiderata* que enumeré para un sistema moral, puedan enumerarse (sin ninguna alteración) como los *desiderata* que debe satisfacer una metodología o un sistema de procedimientos para cualquiera de las principales áreas de los intereses humanos. En la filosofía analítica, los principales intentos por mejorar nuestra comprensión de la naturaleza de la racionalidad han venido de la mano de los filósofos de la ciencia, y esos esfuerzos han dado como resultado dos importantes tendencias.

EL POSITIVISMO LOGICO

En los últimos cincuenta años, la manifestación más clara de la tendencia a pensar que los métodos de «justificación racional» son otorgados por algo así como una lista o *canon* (si bien esos filósofos de la ciencia admiten no haber logrado aún una formalización plena) fue el movimiento conocido como Positivismo Lógico. Los positivistas no sólo confiaban en que los «lógicos de la ciencia» (éste era su término para denominar a los filósofos) lograrían dejar constancia escrita de una descripción (pretendidamente exhaustiva) del método científico, sino que, según ellos, el «método científico» agotaba la propia racionalidad, y la contrastabilidad mediante ese método agotaba la significatividad («el significado de una oración es su método de verificación»); la lista o *canon* determinaría lo que tiene significación cognitiva. Los enunciados contrastables mediante métodos pertenecientes a la lista (los métodos de la matemática, la lógica y las ciencias empíricas) contarían como significativos; los positivistas mantenían que todos los demás eran sólo «pseudo-enunciados», o sinsentidos disfrazados.

Una réplica obvia consistía en afirmar que el criterio de significación del positivismo lógico se *autorrefuta:* el mismo criterio no es ni (a) «analítico» (un término que los positivistas utilizaban para dar cuenta de la lógica y de la matemática), ni (b) empíricamente contrastable. Es bastante extraño que esta crítica produjese en los positivistas lógicos un impacto tan leve que no impidiese el desarrollo de su movimiento. Creo que fue un error no prestar atención a este particular gambito filosófico; el gambito no sólo es correcto, sino que contiene una profunda lección, y una lección que no se limita al positivismo lógico.

El argumento que voy a desarrollar dependerá de la siguiente observación: las formas de «verificación» que los positivistas lógicos autorizaban habían sido *institucionalizadas* por la sociedad moder-

na. Lo que puede ser «verificado», en el sentido positivista, puede ser verificado como correcto (en el sentido no-filosófico o pre-filosófico de «correcto») o como probablemente correcto, o como un éxito científico, según el caso, y el reconocimiento público de la corrección, o de la probable corrección, o del *status* de «teoría científica exitosa», ejemplifica, celebra y refuerza las imágenes del conocimiento y las normas de razonabilidad mantenidas por nuestra propia cultura.

A primera vista, el original paradigma de verificación del positivismo no era un paradigma públicamente institucionalizado. En *Der Logische Aufbau der Welt* de Carnap, la verificación era en última instancia privada, esto es, se basaba en sensaciones cuya cualidad subjetiva o «contenido» se consideraba «incomunicable». Pero Carnap, a instancias de Neurath, convirtió pronto su concepción de la verificación en algo más público, más «intersubjetivo».

Popper ha insistido en la idea de que las predicciones científicas se contrastan con «oraciones básicas», oraciones tales como «El plato derecho de la balanza está más bajo que el izquierdo» que son aceptadas públicamente, pese a que no pueden ser «probadas» a satisfacción del escéptico. Popper ha sido criticado por utilizar en este punto un lenguaje «convencionalista», por hablar como si existiese una convención o una decisión social de aceptar una oración básica; pero creo que lo que nos parece un elemento convencionalista en el pensamiento de Popper es simplemente el reconocimiento de la naturaleza *institucionalizada* de las normas a las que apelamos en los juicios de percepción ordinarios. La naturaleza de nuestra respuesta al escéptico, que nos desafía a «probar» enunciados como «Tengo los pies en el suelo», da cuenta de la existencia de normas sociales que *exigen* estar de acuerdo con tales enunciados en las circunstancias adecuadas.

Wittgenstein sostenía que sin esas normas públicas, normas que son compartidas por un grupo y que constituyen una «forma de vida», no sería posible el lenguaje, ni tampoco el pensamiento. Según Wittgenstein, es absurdo preguntar si la verificación institucionalizada de la que he estado hablando desempeña un papel «realmente» justificativo. En *Sobre la certeza,* Wittgenstein observa que los filósofos pueden proporcionarnos cientos de «justificaciones» epistemológicas del enunciado «Los gatos no brotan en los árboles» —pero ninguna de ellas comienza con algo que sea más seguro (en el sentido institucionalizado de «seguro») que el propio hecho de que los gatos no brotan en los árboles.

Los escépticos no sólo han dudado de los juicios perceptivos, sino también de las inducciones ordinarias. Aunque no estoy siguiendo su distinción entre lo que es *racional* y lo que es *razonable,* Hume hubiera dicho que no hay prueba *racional* de que vaya a nevar en España

este invierno (ni siquiera de que es probable que nieve), aunque hubiese añadido que no sería razonable dudar de que sí que nevará. Con todo, nuestra respuesta al escéptico que nos desafía a que «demostremos» que este invierno nevará en España, da cuenta de la existencia de normas sociales que requieren que se esté de acuerdo con esas «inducciones», del mismo modo que existen tales normas para nuestros juicios perceptivos con respecto a personas cuyos pies están en el suelo o con respecto a la igualdad de los brazos de una balanza.

Las reacciones son algo distintas cuando se llega a las teorías de alto nivel de las ciencias exactas. La gente de la calle no puede «verificar» la teoría especial de la relatividad. De hecho, ni siquiera logra *aprender* la teoría especial en la actualidad, ni tampoco la matemática (relativamente elemental) necesaria para comprenderla, a pesar de que está empezando a ser impartida en los primeros cursos de física de algunos de nuestros colegios. La gente de la calle deja que sean los científicos los que se encarguen de proporcionar una estimación capaz (y socialmente aceptable) de una teoría de este tipo. Dada la inestabilidad de las teorías científicas, no es probable que un científico califique de «verdadera» a *tout court* ni siquiera a una teoría tan exitosa como la relatividad especial. No obstante, la comunidad científica considera que la teoría de la relatividad especial es un «éxito científico» —de hecho constituye un éxito sin precedentes, como la electrodinámica cuántica— que produce «predicciones exitosas» y que «está avalada por un gran número de experimentos». Y son los demás miembros de la sociedad quienes les delegan estos juicios. La diferencia entre este caso y los de las normas institucionalizadas de verificación a los que me he referido antes (dejando a un lado la evanescencia del adjetivo «verdadero») es el *rol* especial desempeñado por los expertos y la deferencia institucionalizada hacia ellos que este caso conlleva; sin embargo, esta circunstancia no es más que un ejemplo de la división del trabajo intelectual (por no mencionar las relaciones de autoridad intelectual). Son las autoridades nombradas por la sociedad, cuya autoridad se reconoce mediante multitud de prácticas y ceremonias, quienes juzgan que la relatividad especial y la electrodinámica cuántica son las «mejores teorías científicas que tenemos» y es en este sentido en el que dicho juicio está institucionalizado.

Recientemente, se me ocurrió la idea de que Wittgenstein posiblemente considerase que *sólo* pueden ser completamente verdaderos (o correctos o adecuados o justificados) aquellos enunciados que pueden ser verificados de alguna forma «institucionalizada». No pretendo sugerir que algún filósofo sostuviese alguna vez que *todo* aquello que en nuestra sociedad cuenta como «justificación» sea realmente tal. Por lo general, los filósofos diferencian entre instituciones constitutivas de nuestros conceptos e instituciones que tienen algún otro *status,* si bien existe cierta controversia con respecto a cómo realizar

tal distinción. Sugiero que Wittgenstein consideraba que lo que puede decirse correctamente o no en los «juegos de lenguaje» en los que nos hallamos inmersos viene determinado por algún subconjunto de nuestras normas institucionalizadas de verificación, y que no existe corrección o incorrección objetiva más allá de éste. Pese a que tal interpretación se ajusta mucho a lo que Wittgenstein dice —por ejemplo, a su énfasis en la necesidad de «un acuerdo en nuestros juicios», hasta para poseer siquiera un concepto— no estoy convencido de que sea correcto. Cuando Wittgenstein habla de «nuestros juicios» es bastante vago con respecto a quién constituye ese «nosotros». Ignoro si sus «formas de vida» corresponden a las normas institucionalizadas que acabo de mencionar. Esta interpretación se me ocurrió al leer *Lectures and Conversations* de Wittgenstein. En este libro, Wittgenstein rechaza tanto el psicoanálisis como la teoría de la evolución de Darwin (aunque, a diferencia de los positivistas, no consideraba que el lenguaje psicoanalítico *careciese de significado;* además, sentía admiración por el «ingenio» de Freud). Lo que Wittgenstein opinaba sobre el psicoanálisis (al que llama «mito») no tiene demasidado interés, ya que mucha gente considera —en mi opinión, equivocadamente— que el psicoanálisis es más o menos un sinsentido. Pero su rechazo de la *evolución* es completamente sorprendente[2]. Wittgenstein contrasta la teoría de Darwin con las teorías físicas, siendo el resultado desfavorable a la primera. («Una de las cosas más importantes con respecto a la explicación es que ha de funcionar, que ha de permitirnos predecir. La física está relacionada con la ingeniería. El puente debe aguantar», *Lectures on Aesthetics,* p. 25.) Y afirmaba que la gente se dejó persuadir «sobre una base extremadamente endeble». «Si en

---

[2] Lo que Wittgenstein afirmaba con respecto a la evolución era que «la gente llegó a convencerse sobre una base extremadamente endeble. No hubo ninguna actitud que afirmase: "No sé. Es una hipótesis interesante que eventualmente puede ser confirmada"», *Lectures on Aesthetics,* p. 26, en Cyril Burret (ed.), *L. Wittgenstein: Lectures and Conversations,* Berkeley University of California Press, 1964. Wittgenstein no dice qué constituiría una «confirmación» de la teoría de la evolución, pero el parágrafo sugiere que lo que tenía en mente era la observación efectiva del proceso de formación de especies. («¿Vio alguien cómo se producía este proceso? No. ¿Ha habido alguien que lo haya visto ocurriendo en el presente? No. La evidencia de la reproducción es sólo una gota en un cubo».) Es instructivo comparar la actitud de Wittgenstein con la de Monod: «Como el propio Darwin declaró, la teoría de la evolución requería el descubrimento de la genética mendeliana, el cual se llevó a cabo, naturalmente. Este es un ejemplo, y un ejemplo importante, de lo que se entiende por contenido de una teoría, por contenido de una idea ... (una) buena teoría o una buena idea será mucho más fértil y extensa de lo que su inventor podía intuir en su tiempo. A medida que va dando de sí, la teoría puede juzgarse con precisión según este tipo de desarrollo, aun cuando gran parte de lo que de ella procederá no pueda predecirse». [J. MONOD: «On the Molecular Theory of Evolution», en Harre, R. (ed.), *Problems of Scientific Revolution: Progress and Obstacles to Progress in the Sciences,* Oxford, 1975.]

último término usted olvida por completo toda cuestión verificativa, acabará francamente convencido de que las cosas debieron suceder así».

De nuevo, los grandes debates en torno a la «analiticidad» que continuaron en la década de los cincuenta guardan relación con el deseo de los filósofos de encontrar un fundamento objetivo e *incontrovertible* para sus argumentos. La «analiticidad», es decir, la doctrina de la verdad en virtud únicamente del significado, fue impugnada debido a su *uso abusivo* por parte de estos últimos. Pero ¿por qué era tan seductor para los filósofos proclamar que muchas cosas que son, en un sentido *no*-inteligible, «reglas del lenguaje», o consecuencias de las reglas del lenguaje, eran analíticas o «conceptualmente necesarias» o algo parecido? Creo que la respuesta es que los filósofos pensaban que la idea de que existe un conjunto definido de *reglas del lenguaje* y que *éstas* pueden establecer lo que es y lo que no es racional, tenía dos ventajas: (1) las «reglas del lenguaje» son prácticas constitutivas e institucionalizadas (o normas que subyacen a tales prácticas) y, como tales, tienen el *status* «público» que he descrito. (2) Se afirmaba al mismo tiempo que sólo los filósofos (y no los lingüistas) podían descubrir esas cosas tan misteriosas. Fue una buena idea mientras duró, pero estaba condenada a ser refutada, y lo fue.

Llamaré concepción *criterial* de la racionalidad a cualquier concepción de acuerdo con la cual la aceptabilidad racional se defina mediante normas institucionalizadas. Los positivistas lógicos, Wittgenstein (al menos en la interpretación que he ensayado y que he reconocido como dudosa) y algunos filósofos [3] del «lenguaje ordinario» de

[3] Sería posible desarrollar una filosofía del «lenguaje ordinario» que no se comprometiese con la verificación pública y «criterial» de las tesis filosóficas, si fuese posible desarrollar y avalar una concepción en la que las normas que gobiernan las prácticas lingüísticas no puedan descubrirse mediante una investigación empírica ordinaria. En *Must We Mean What We Say,* Stanley Cavell dió un paso significativo en esta dirección, argumentando que esas normas pueden conocerse mediante una especie de «autoconocimiento», autoconocimiento que compara con el discernimiento que se alcanza a través de la terapia y también con el conocimiento transcendental buscado por la fenomenología. Pese a estar de acuerdo con Cavell en que mi conocimiento como hablante nativo de que ciertos usos se desvían o no, no es un conocimiento inductivo «externo» —no necesito evidencia para saber que en mi dialecto del castellano se dice «roto» y no «rompido»— me inclino a pensar que este acceso privilegiado del hablante no alcanza a las *generalizaciones* sobre la corrección y la incorrección. Si afirmo (como Cavell) que es parte de la regla para el uso correcto de expresiones de la forma *X es intencionado* el que en *X* haya algo «sospechoso», entonces estoy anticipando una *teoría* con el fin de explicar mis intuiciones con respecto a casos específicos, y no dando mera cuenta de éstas. Es cierto que en psicoterapia también ocurre algo parecido; pero no me inclino a conceder al autoconocimiento ningún tipo de inmunidad frente a la crítica de los demás, incluyendo aquellas críticas que dependen del ofrecimiento de explicaciones rivales, en uno u otro caso. Si se concede legitimidad a tales críticas, la actividad de descubrir tales normas comienza a parecerse a la ciencia social o a la

Oxford, aunque no todos, compartían una concepción criterial de la racionalidad —aun cuando discrepaban en otros puntos, tales como si se debe llamar «carentes de significado» a los enunciados inverificables, o con respecto a si algunas proposiciones éticas podían ser «conceptualmente necesarias».

El gambito al que me refería al comienzo, el gambito que refuta el principio de verificación del positivismo lógico, es *incisivo* precisamente porque refuta cualquier argumento a favor de una concepción criterial de la racionalidad, esto es, porque refuta la tesis de que nada es racionalmente aceptable excepto si es criterialmente verificable.

Es significativo que la forma de hablar de los filósofos que acabo de mencionar sugiera que sus argumentos tienen el mismo tipo de *finalidad* que una demostración matemática o que una prueba experimental en física, y que, aunque los positivistas llamaron a su trabajo *lógica* de la ciencia, aunque los wittgensteinianos exhibieron una increíble arrogancia frente a aquellos filósofos que no podían «ver» que toda actividad filosófica de tipo pre-wittgensteiniano o no-wittgensteiniano carecía de sentido, y aunque los filósofos del lenguaje ordinario calificaron de «faltas garrafales» a los argumentos de sus oponentes, estuviesen en su bando o en el de los filósofos del lenguaje no-ordinario (como si los errores filosóficos se pareciesen a las faltas de una prueba aritmética) no existe ninguna posición filosófica de importancia que pueda ser verificada de la forma conclusiva y culturalmente reconocida que he descrito. En resumen, si es cierto que sólo pueden ser racionalmente aceptables aquellos enunciados que pueden verificarse criterialmente, este mismo enunciado no puede ser verificado criterialmente y, *por tanto, no es racionalmente aceptable.* Si es que existe tal cosa como la racionalidad —y nos comprometemos a creer en *alguna* noción de racionalidad al tomar parte en las actividades de *hablar y argumentar*— entonces la actividad de *argumentar* en favor de una posición que la identifica a (o la convierte en un subjconjunto propio de) lo que las normas institucionalizadas determinan ya como instancias de racionalidad, es una actividad que se autorrefuta. Pues esas normas no pueden garantizar por sí solas la corrección, o la probable corrección, de ningún argumento de esta índole.

No estoy negando la posibilidad de argumentación y justificación

---

historia —áreas en las que (como he indicado) las descripciones tradicionales de «Método Científico» arrojan poca luz (véase mi libro *Meaning and The Moral Sciences,* Routledge and Kegan Paul, 1978).

En cualquier caso, y sea cual sea su *status,* no veo ninguna razón para creer que las normas de uso del *lenguaje* decidan la extensión de predicados como «racionalmente aceptable», «justificado», «bien confirmado», y otros parecidos.

racional en filosofía; más bien me he visto obligado a reconocer algo que probablemente es evidente para los legos, aunque no para los filósofos, a saber: no podemos apelar a normas *públicas* para decidir qué enunciados filosóficos son racionalmente argumentables y justificables. La afirmación que aún hoy es frecuente oír, la afirmación de que la filosofía es «análisis conceptual», que los *propios conceptos* determinan qué argumentos filosóficos son correctos, es (cuando se combina con la doctrina de que los conceptos son normas o reglas que subyacen a la prácticas lingüísticas *públicas*) tan sólo una forma encubierta de afirmar que toda justificación racional en filosofía es criterial, y que la verdad filosófica es (si exceptuamos los «errores garrafales») *públicamente demostrable* en la misma medida en que lo es la verdad científica. Tal opinión me parece poco razonable a la luz de toda la historia del problema, incluyendo la historia reciente.

Los argumentos sobre la religión y sobre la ideología secular corren la misma suerte que los argumentos filosóficos. Una argumentación entre un inteligente marxista y un inteligente liberal tendrá, a fin de cuentas, el mismo carácter que una disputa filosófica, si bien existen más hechos empíricos relevantes. Todos tenemos puntos de vista en religión, política o filosofía, y argumentamos en favor de éstos, criticando los de los demás. En realidad, poseemos argumentos con idéntico carácter en historia, sociología y psicología clínica, e incluso en la «ciencia», dejando aparte las ciencias exactas. Es verdad que los positivistas lógicos ampliaron su descripción del «método científico» para que *incluyese* estas materias. Pero una vez que su descripción se amplía hasta ese punto, no pueden demostrar que excluya algo con claridad, sea lo que sea. (V. capítulo 8.)

Recordemos que los positivistas *concedían* que el principio de verificación carecía de «significado cognitivo». Afirmaban que era una *propuesta,* y como tal, no era ni verdadera ni falsa. No obstante, sí que argumentaban a favor de su propuesta y sus argumentos no eran (y no tenían por qué ser) puntos de partida[4]. Así que el problema permanecía irresuelto.

---

[4] El argumento más débil que se ofreció en defensa de la interpretación del principio de verificación como una propuesta consistía en afirmar que «explicaba» la noción «pre-analítica» de significación plena. (Esta afirmación se discute en «How Not to Talk About Meaning», en mi libro *Mind, Language and Reality, Philosophical Papers,* vol. 2, Cambridge University Press, 1975.) Reichenbach defendió una forma del principio de verificación (en *Experience and Prediction*) según la cual éste *preservaba todas aquellas diferencias de significado relevantes para la conducta.* Ante una objeción obvia (la creencia no empírica en una divinidad —Reichenbach emplea el ejemplo de los egipcios que adoran a los gatos— puede alterar la conducta), Reichenbach replica proponiendo interpretar la oración «Los gatos son animales sagrados» como «Los gatos inspiran te-

En resumidas cuentas, lo que hicieron los positivistas lógicos y Wittgenstein (y quizá también el último Quine) fue *producir filosofías que no dejaban lugar a la actividad racional de la filosofía.* Sus concepciones se autorrefutan por este motivo, pero también porque el pequeño gambito del que he estado hablando representa un argumento significativo, del tipo que los filósofos denominan «argumento trascendental»: argumentar con respecto a la naturaleza de la racionalidad (la tarea *par excellence* de los filósofos) es una actividad que presupone una noción de justificación más amplia que la positivista, y, en realidad, más amplia que la noción de racionalidad criterial institucionalizada.

EL ANARQUISMO SE AUTORREFUTA

A continuación discutiré otra tendencia filosófica muy distinta de la anterior. La obra de Kuhn *La estructura de las revoluciones científicas (ERC)* cautivó a gran número de lectores, pero horrorizó a la mayoría de los filósofos de la ciencia, debido a su énfasis en lo que parecen constituir los *determinantes irracionales* de la aceptación de una teoría científica y a su uso de términos como «conversión» y «cambio gestáltico». De hecho, Kuhn puso de manifiesto bastantes aspectos importantes acerca de las teorías científicas y acerca de cómo debe concebirse la actividad científica. Ya he expresado mi acuerdo en relación con la relevancia de las nociones de *paradigma, ciencia normal y revolución científica,* en otro lugar. Llegado este punto, quiero dirigir mi atención hacia aquellos aspectos del libro de Kuhn con los que no simpatizo, lo que en otro sitio he descrito como «el relativismo extremo de Kuhn».

Según la lectura de Kuhn que cautiva más estudiantes de segundo curso, Kuhn afirma que en la ciencia no hay justificación *racional,* sino tan sólo cambios gestálticos y conversiones. Kuhn ha rechazado esta interpretación de *ERC* para introducir posteriormente la noción

mor y respeto a los que los adoran». ¡Está claro que, en el caso de alguien que adore a los gatos, la aceptación de este sustituto sí que alteraría su conducta!

El punto de vista de Carnap es más interesante. De acuerdo con Carnap, todas las reconstrucciones racionales son propuestas. Las únicas cuestiones fácticas tienen que ver con las consecuencias lógicas y empíricas de aceptar tal o cual reconstrucción. (Carnap comparó la «elección» de una reconstrucción racional con la elección de un motor para un avión.) Y extrajo la conclusión de que en filosofía hemos de ser tolerantes con respecto a reconstrucciones racionales divergentes. Sin embargo, este principio de tolerancia (como Carnap lo llamaba) *presupone* el principio de verificación. Pues la doctrina según la cual ninguna reconstrucción racional es particularmente correcta, ni corresponde a cómo «son realmente» las cosas, la doctrina de que todas las «cuestiones externas» carecen de sentido cognitivo, es precisamente el principio de verificación. Y aplicar el principio de tolerancia al propio principio de verificación sería argumentar en círculo.

de «racionalidad no-paradigmática», la cual, si no es precisamente la misma noción que yo denomino «racionalidad no-criterial», está íntimamente relacionada con ella.

La tendencia que la mayor parte de los lectores creyeron detectar en *ERC* se manifiesta, sin ninguna duda, en *Contra el método* de Paul Feyerabend. Feyerabend, como Kuhn, puso énfasis en que diferentes culturas y diferentes épocas históricas producen diferentes paradigmas de racionalidad. Sugiere que *nuestras* concepciones de la racionalidad están en buena medida determinadas por lo que *nosotros* denominaríamos «lo irracional». En efecto, pese a que no lo dice así, Feyerabend sugiere que la concepción moderna (científico-tecnológica) de la racionalidad es fraudulenta en su propia base. (Creo detectar una tendencia similar en Michel Foucault.) Y va más allá de Kuhn o de Foucault al sugerir que incluso la tan cacareada superioridad instrumental de nuestra ciencia es algo así como un trampa. Los curanderos pueden hacer más para aliviar los dolores que los médicos, viene a decir Feyerabend.

No quiero discutir estas afirmaciones tan terriblemente radicales, si bien fueron éstas las que decidieron a Feyerabend a llamar a su posición «anarquismo». Deseo discutir una afirmación que Kuhn establece tanto en *ERC* como en artículos subsiguientes, y que Feyerabend realiza tanto en *Contra el método* como en artículos más técnicos. Esta afirmación constituye la tesis de la *inconmensurabilidad.* Quiero afirmar que esta tesis se autorrefuta, al igual que las tesis positivas con respecto al significado y a la verificación. En resumen, quiero afirmar que *las dos* filosofías de la ciencia más influyentes del siglo veinte (sin duda alguna, las dos que más han interesado a los científicos y a los no-filósofos, y las dos únicas de las que probablemente haya oído hablar el lector medianamente culto), se autorrefutan. Como filósofo de la ciencia encuentro algo problemático que esto sea así, por supuesto. ¿Cómo se explica esta situación? Nos ocuparemos en breve de este problema.

La tesis de la inconmensurabilidad afirma que los términos utilizados por otra cultura (pongamos por caso el término «temperatura», tal como lo utilizó un científico del siglo diecisiete) no pueden hacerse equivalentes en significado o referencia con ninguno de *nuestros* términos o expresiones. Como dijo Kuhn, los científicos que se hallan bajo paradigmas diferentes habitan «mundos diferentes». La palabra «electrón», tal como se utilizó alrededor de 1900, se refería a objetos de un «mundo». Pero tal como se utiliza hoy, se refiere a objetos de un «mundo» completamente distinto. Se supone que esta tesis se aplica tanto al lenguaje observacional como al denominado

«lenguaje teórico»; en realidad, de acuerdo con Feyerabend, el lenguaje ordinario *es* tan sólo una teoría falsa.

En esta ocasión la réplica consiste en que, si la tesis fuera realmente verdadera, no habría manera de traducir otros lenguajes —ni siquiera estadios anteriores de nuestro propio lenguaje. Y si nos es absolutamente imposible interpretar los ruidos de los organismos, dejamos de tener razones para considerarlos como seres que *piensan y hablan,* ni siquiera para considerarlos *personas.* En resumen, si Feyerabend (y Kuhn, en sus tesis más radicales con respecto a la inconmensurabilidad) estuviera en lo cierto, sólo podríamos caracterizar a los miembros de otras culturas, incluyendo a los científicos del siglo diecisiete, como animales que producen respuestas a estímulos (incluyendo entre éstos a aquellos ruidos que, curiosamente, se parecen al inglés o al italiano). Decir que Galileo poseía nociones que son «inconmensurables» con las nuestras, *para seguidamente describirlas con detalle,* es algo totalmente incoherente.

Smart [5] plantea este problema en un indulgente ensayo sobre Feyerabend:

> Sin duda, para observar Mercurio no importa que apuntemos el telescopio hacia la copa de un árbol, digamos, y no, como predijo la teoría newtoniana, hacia el cimborrio de una chimenea. Y, seguramente, podemos hablar de árboles, cimborrios de chimeneas y de telescopios, independientemente de la elección entre la teoría newtoniana y la einsteniana. Feyerabend podría reconocer, sin embargo, que usamos geometría euclidiana y óptica no-relativista para la teoría de nuestro telescopio. Afirmaría que ésta no es la verdad real con respecto a nuestro telescopio, al árbol y al cimborrio de la chimenea, pero que, no obstante, es legítimo pensar de esta manera de cara a discutir las pruebas observacionales de la teoría de la relatividad general, ya que tenemos bases teóricas suficientes para saber que nuestras predicciones no se verán afectadas (hasta el límite del error observacional) si nos beneficiamos de esta ventaja en el cálculo.

Pero la maniobra de rescate de Smart se topa con el problema de que, hasta para decir que las «predicciones» son las mismas, debo comprender algo del lenguaje no-relativista euclidiano. Si todas *las palabras tienen una significación distinta,* ¿en qué sentido pueden permanecer «inalteradas» nuestras predicciones? ¿Cómo puedo siquiera traducir las partículas lógicas (las palabras «si-entonces», «no», etc.), al italiano del siglo diecisiete, si no puedo encontrar un manual de traducción que asocie el italiano del siglo diecisiete y el castellano moderno, y que elabore algún tipo de sentido sistemático del corpus del

[5] J. J. C. SMART, «Conflicting Views about Explanation», en R. Cohen y M. Wartofski (eds.), *Boston Studies in The Philosophy of Science, Volume II: in Honor of Philipp Frank,* New York, Humanities Press, Inc., 1965.

siglo diecisiete, tanto en sí mismo como en su marco extralingüístico? Aun cuando yo sea un hablante que emplee ambas teorías (como imagina Smart), ¿tengo algún motivo para establecer una equivalencia entre alguna palabra de mi teoría newtoniana y alguna otra de mi teoría de la relatividad general?

El punto que intento hacer ver adquiere más contraste si cabe al aplicarle alguna de las observaciones de Quine y Davidson con respecto al significado y a la práctica de la traducción. Una vez se acepta que podemos hallar un manual de traducción que «funcione» en el caso de un texto del siglo diecisiete, al menos en el contexto fijado por nuestros intereses y por el uso que vayamos a hacer de la traducción, ¿qué sentido tiene decir, *en este contexto,* que la traducción no capta «realmente» el sentido o la referencia del original? Después de todo, parece como si se afirmase que tenemos (o que es posible obtener) criterios de mismidad de sentido o de referencia, dejando a un lado nuestros esquemas de traducción y nuestros requisitos explícitos o implícitos para la adecuación empírica de tales esquemas. Afirmar que una traducción no capta con exactitud el sentido o la referencia del original puede ser entendida como un reconocimiento de que es posible encontrar un esquema de traducción mejor, pero decir que ningún esquema de traducción capta el sentido o la referencia «reales» produce sólo la ilusión de un sentido. La sinonimia existe sólo como una relación, o, mejor dicho, como una familia de relaciones (todas ellas algo vagas) que empleamos para establecer la equivalencia entre expresiones diferentes, según los propósitos de la interpretación. La idea de que existe algo como la sinonimia «real» dejando aparte todos los procedimientos factibles de interpretación, ha sido descartada como un mito.

Supongamos que alguien nos dice que la palabra alemana «Rad» puede traducirse como «rueda». Si pasa luego a afirmar que su traducción no es perfecta, lo que normalmente esperamos es que nos indique cómo podría ser mejorada, complementándola con una glosa, o de cualquier otro modo. Pero si continúa diciendo que «Rad» puede traducirse como «rueda» pero que no se refiere «realmente» a ruedas, ni a ningún otro objeto reconocido como tal en nuestro esquema conceptual, ¿sacaremos algún provecho de su afirmación? Decir que *A* puede traducirse por «rueda» es decir que, en la medida en que sea posible fiarse de la traducción, *A se refiere* a ruedas.

La razón por la cual la tesis de la inconmensurabilidad intriga tanto a la gente (aparte del atractivo que parecen tener todas las ideas incoherentes) quizá sea la tendencia a fundir o confundir concepto y concepción. Esta distinción es borrosa en la misma medida en que lo es la distinción analítico-sintético. Pero está involucrada en toda interpretación, si bien es relativa a cada interpretación. Cuando traducimos una palabra como «temperatura», por ejemplo, establecemos la

equivalencia referencial, y también de sentido (en la medida en que nuestra traducción sea fiel), con la expresión traducida mediante nuestro término «temperatura», al menos como lo usamos en ese contexto. (Naturalmente, podemos utilizar varios recursos —como glosas especiales— para delimitar o delinear cómo utilizamos la palabra «temperatura» o cualquier otra en un determinado contexto.) En este sentido, establecemos la equivalencia del «concepto» en cuestión con nuestro propio «concepto» de «temperatura». Pero esta equivalencia no es incompatible con el hecho de que los científicos del siglo diecisiete, o quien se nos antoje, tuviesen probablemente una diferente *concepción* de la temperatura, esto es, un conjunto de creencias con respecto a la temperatura y a su naturaleza que fuese distinto del nuestro, diferentes «imágenes del conocimiento», y también diferentes creencias últimas con respecto a otras muchas cosas. Que las concepciones difieran no demuestra la imposibilidad de traducir ninguna concepción «de un modo realmente correcto», como a veces se supone; por el contrario, no podríamos decir que nuestras concepciones difieren, y en qué difieren, si no pudiésemos traducirlas.

Pero si, a fin de cuentas, las concepciones resultan ser diferentes, ¿cómo podemos llegar a saber si un esquema de traducción es «válido»? Desde Vico hasta nuestros días, diversos pensadores han respondido a esta pregunta afirmando que el éxito interpretativo no requiere que las creencias de las personas a las que se está traduciendo resulten ser nuestras *mismas* creencias, sino que nos resulten *inteligibles*. Esta es la base de las diversas máximas de caridad interpretativa o del «beneficio de la duda», tales como «Interprételos de forma que se revelen como personas que creen cosas verdaderas y que son amantes del bien» o «Interprételos de forma que sus creencias resulten razonables a la luz de lo que les han enseñado y de lo que han experimentado», o la propia instrucción de Vico de maximizar la *humanidad* de la persona que se está interpretando. Es un hecho constitutivo acerca de la experiencia humana, en un mundo en el que diferentes culturas interactúan a lo largo de la historia —aunque individualmente sufran cambios más lentos o más rápidos que los nuestros— que somos capaces —y es cuestión de experiencia humana universal— de interpretar las creencias, los deseos y las proferencias de los demás de forma que todas tengan algún tipo de *sentido*.

No es sorprendente que Kuhn y Feyerabend rechacen la idea de *convergencia* en el conocimiento científico. Puesto que no hablamos de las mismas cosas de las que hablaban los científicos anteriores, no estamos obteniendo un conocimiento cada vez mayor acerca de los mismos objetos macro o microscópicos. Kuhn arguye que el «progreso» de la ciencia es únicamente instrumental; estamos mejor dotados para transportar a la gente de un lugar a otro, etc. Pero esto también

resulta incoherente. A menos que locuciones como «transportar a la gente de un lugar a otro» conserven algún grado de fijación en la referencia, ¿cómo podemos comprender la noción de éxito instrumental de un modo estable?

El argumento que acabo de emplear está relacionado esencialmente con los conocidos argumentos de Kant con respecto a las precondiciones del conocimiento empírico. Respondiendo a la aserción de que el futuro podría carecer de leyes por completo, podría desbaratar cualquier «inducción» que realizásemos, Kant señaló que, si hay algún futuro —algún futuro para *nosotros* de todos modos, algún futuro que podamos aprehender y conceptualizar como seres pensantes de cara a decir si nuestras predicciones son verdaderas o falsas—, entonces, de hecho, *no* deben haber sido violadas muchas regularidades. Si no, ¿por qué llamarlo *futuro*? Por ejemplo, cuando imaginamos bolas saliendo de una en una en algún orden irregular, olvidamos que *para poder siquiera decir que son bolas, o en qué orden salieron,* tenemos que atenernos a muchas regularidades. La *comparación* presupone la existencia de algunas conmensurabilidades.

La baza que Kuhn y Feyerabend podrían jugar en respuesta a todas estas críticas, si bien no se sentirían muy felices llevándola a cabo, sería introducir algún tipo de dicotomía observacional/teorético. Podrían aceptar la conmensurabilidad, la traducibilidad e incluso la convergencia con respecto a los hechos observacionales, restringiendo la tesis de la inconmensurabilidad al vocabulario teóretico. Y aun así habría problemas (¿por qué no describir los significados de los términos teoréticos *mediante* sus relaciones con el vocabulario observacional, *à la Ramsey*?). Pero Kuhn y Feyerabend rechazan esta alternativa, y con razón, pues, de hecho, los principios de caridad interpretativa son tan necesarios en el «lenguaje observacional» como en el «lenguaje teorético». Consideremos, por ejemplo, una palabra corriente, como «hierba». Diferentes hablantes tendrán diferentes estereotipos de hierba (la hierba tiene diferentes formas y colores en diferentes lugares) y diferentes concepciones de la hierba, dependiendo de cuándo y dónde vivan. Aun cuando todos los hablantes deban saber que la hierba es una planta, so pena de reconocer que tienen un concepto completamente distinto, la concepción de «*planta*» supone hoy la *fotosíntesis,* y la de hace dos siglos no. Sin la caridad interpretativa, que nos conmina a establecer la equivalencia (al menos en algunos contextos ordinarios) entre el término «planta» de hace doscientos años y el término «planta» de nuestros días, y entre el término «hierba» de hace doscientos años y el término «hierba» de nuestros días, no podría hacerse ningún enunciado acerca de la referencia de este término hace doscientos años. No sólo es la interpretación de los términos de géneros naturales la que depende en tal grado de los principios de caridad: el término artificial «pan» planteará los mis-

mos problemas. En realidad, sin la caridad interpretativa, ni siquiera sería posible establecer la equivalencia de un simple término de color, como «rojo», utilizado por diferentes hablantes. Siempre interpretamos el discurso como un todo, y la interpretación de los términos de «observación» depende de la interpretación de los términos «teoréticos» en la misma medida que la de los segundos depende de la de los primeros.

He ofrecido, una vez más, un argumento trascendental. Nuestras concepciones fundamentales nos comprometen a tratar como personas no sólo a nuestras porciones espacio-temporales presentes, sino también a nuestros propios pasados, a nuestros ancestros, y a los miembros de las demás culturas, tanto pasadas como presentes, y como he argumentado, esto significa atribuirles referencias y conceptos compartidos, por muy diferentes que sean las *concepciones* que les atribuyamos. Y no sólo compartimos con los demás objetos y conceptos, en la medida que tenga éxito el ejercicio de la interpretación, sino también concepciones de lo razonable, de lo natural, etc. Pues recordemos que lo que justifica un esquema de interpretación es que reproduce la conducta de los demás haciéndola al menos mínimanente razonable según *nuestros* conocimientos. Por muy diferentes que sean nuestras imágenes del conocimiento y nuestras concepciones de la racionalidad, compartimos un vasto fondo de suposiciones y creencias acerca de lo que es razonable incluso con la cultura más rara que consigamos interpretar.

¿POR QUE ES INCONSISTENTE EL RELATIVISMO?

Es un tópico entre los filósofos que el relativismo (total) es inconsistente. Después de todo, ¿no es *obviamente* contradictorio *mantener* un punto de vista mientras se mantiene al mismo tiempo que *ningún* punto de vista es correcto, ni está más justificado que cualquier otro? Alan Garfinkel ha puesto de manifiesto el problema muy mordazmente. Conversando con sus estudiantes californianos y remedando su modo de hablar, dijo en una ocasión: «Usted puede no estar de vuelta de donde yo ya estoy de vuelta, pero yo sé que el relativismo no es *verdadero para mí*»... Si cualquier punto de vista es tan bueno como cualquier otro, ¿por qué no puede ser *el punto de vista que afirma que el relativismo es falso* un punto de vista tan bueno como cualquier otro?

La plétora de doctrinas relativistas que hoy día están a la venta (lanzadas al mercado por pensadores sumamente inteligentes) indica que esta simple refutación no bastará. ¿Por qué un relativista inteligente habría de reconocer que todas las opiniones son igualmente verdaderas *(para él)?* El relativista no puede impedir que usted (o Alan

Garfinkel) afirme que su opinión no es *verdadera para usted* (o no tiene justificación, etc.): pero si conserva su presencia de ánimo, puede espetarle que lo que es verdad para usted dista de ser tan relevante *(para él)* como lo que es verdad para él. Después de todo, ¿hay algún concepto de algo que sea más relevante que el que uno tiene? ¿Es *realmente* inconsistente considerar *verdad, justificado,* etc., como nociones *relativas*?

La respuesta afirmativa requiere un argumento más elaborado que el argumento unilineal (y, no obstante, muy acertado) aducido por Garfinkel. Lo importante es darse cuenta de que si todo es relativo, lo relativo también es relativo. Mas esto requiere alguna explicación.

Quizá fue Platón el primer filósofo en emplear (contra Protágoras) el tipo de argumento que tengo en mente. Protágoras (aparentemente, un relativista impenitente) afirmaba que cuando digo *X,* tendría que decir «Yo creo que *X*». De modo que Protágoras afirmaría que cuando digo «La nieve es blanca» tendría que decir que Hilary Putnam cree que la nieve es blanca, y que lo que Robert Nozick quiere decir con la misma proferencia es que Robert Nozick cree que la nieve es blanca. Una exposición más sofisticada de la misma idea sostendría que cuando digo «La nieve es blanca» estoy utilizando esta proferencia para afirmar que es verdad-para-mí que *la nieve es blanca,* mientras que cuando Robert Nozick dice las mismas palabras normalmente afirma que es verdad-para-él que *la nieve es blanca* (o, al menos, sólo consideraría correcto su enunciado en el caso en que fuera verdadero-para-*él*). Se sigue (de acuerdo con Protágoras) que ninguna proferencia tiene el mismo *significado* para mí y para cualquier otra persona; como hemos visto antes, hay una íntima conexión entre relativismo e inconmensurabilidad. El contra-argumento de Platón consistía en que si cada enunciado *X* significa «Creo que *X*», entonces (siguiendo a Protágoras) debo decir realmente,

(1) Yo creo que yo creo que la nieve es blanca.

Pero el procedimiento de añadir «Yo creo» puede iterarse infinitamente. De acuerdo con Protágoras, el significado último de «La nieve es blanca» no es entonces (1) sino

(2) Yo creo que yo creo que yo creo... (con número infinito de «yo creo») que la nieve es blanca.

Platón consideró esto como una *reductio ad absurdum.* Sin embargo, el argumento de Platón, tal como está, no es un buen argumento. ¿Por qué Protágoras no habría de estar de acuerdo en que su análisis se aplicase a sí mismo? No se sigue que el argumento *deba* autoaplicarse un número o infinito de veces, sino sólo que *puede* autoaplicarse *algún número finito de veces.* Pero Platón se había dado cuenta de algo muy grave.

Cuando uno se topa por primera vez con el relativismo, la idea *parece* bastante simple. La idea, en una primera formulación, es que cada persona (o, en una moderna formulación *sociológica,* cada cultura, o, a veces, cada «discurso») tiene sus propios puntos de vista, estándares, presuposiciones, y que la verdad (y también la justificación) es relativa a éstos. Se da por sentado, desde luego, que el que *X* sea verdadero (o esté justificado) relativamente a éstos es *en sí mismo* algo «absoluto».

Los modernos estructuralistas, como Foucault, dan a entender que la justificación *relativa a un discurso* es en sí misma algo absoluto —es decir, de ningún modo relativa. Pero si los enunciados de la forma «*X* es verdadero (está justificado) relativamente a la persona *P*» son en sí mismos *absolutamente* verdaderos o falsos, entonces, después de todo, existe una noción absoluta de verdad (o de justificación) y no sólo la verdad-para-mí, verdad-para-Nozick, verdad-para-usted, etc. Un relativista *total* debería decir que *X es verdadero relativamente a P es en sí mismo relativo.* Llegados a este punto, comienza a tambalearse nuestra comprensión de lo que significa la postura relativista, como Platón observó.

La estrategia platónica contra el relativismo no parece haberse ultimado hasta hace bien poco. Wittgenstein la fortaleció brillantemente, sobre todo en el argumento del lenguaje privado (al que ya aludimos en el capítulo 3).

Muchos exegetas leen el argumento del lenguaje privado simplemente como un argumento contra la teoría de la «verdad-copia». Y sin duda alguna, la brillante demostración wittgensteiniana de que la teoría de la referencia-similitud no resuelve ni siquiera el problema de la referencia a sensaciones es parte del ataque librado contra el realismo metafísico. Aun así, prefiero leer el argumento desdoblándolo en un par de argumentos absolutamente tradicionales (al menos Kant habría aprobado ambos) contra *dos* posiciones: la posición realista y la posición relativista. El intento de leer todo el argumento como un argumento exclusivamente antirrealista lo convierte en algo bastante artificial.

A Wittgenstein le interesaba atacar una forma de relativismo conocida como «solipsismo metodológico». Un «solipsista metodológico» es un no-realista o «verificacionista» que está de acuerdo en que la verdad debe concebirse emparentada de alguna forma con la aceptabilidad racional, pero que sostiene que toda justificación viene dada en última instancia en términos de experiencias de las que cada uno tiene un conocimiento *privado.* De este modo, yo tengo *mi* conocimiento de cuáles de *mis* experiencias verificarían que la nieve es blanca, y Bob Nozick tiene *su* conocimiento de cuáles de *sus* experiencias verificarían que la nieve es blanca: cada enunciado tiene un sentido diferente para cada sujeto pensante.

El argumento de Wittgenstein me parece un argumento excelente contra el relativismo en general. Consiste en que, en última instancia, el relativista no puede dar ningún sentido a la distinción entre *estar en lo cierto y creer que se está en lo cierto,* y esto significa que, al fin y al cabo, no hay diferencia alguna entre *afirmar o pensar,* por una parte, y *producir ruidos (o imágenes mentales)* por otra. Pero esto significa que (según esta concepción) yo no soy un *ser que piensa,* sino un mero animal. Mantener este punto de vista nos compromete con una especie de suicidio mental.

Para ver que Wittgenstein estaba en lo cierto, consideremos algo que él no tuvo en cuenta, esto es, de qué manera podría el relativista *intentar* trazar la distinción que Wittgenstein le niega, la distinción entre estar en lo cierto y creer que se está en lo cierto.

El relativista podría adoptar la idea de que la verdad es una *idealización* de la aceptabilidad racional. Podría mantener que *X* es verdadero-para-mí si «*X tiene justificación para mí*» *fuese* verdadero siempre que lo examinase con suficiente cuidado, razonase lo suficiente, etc. Pero los condicionales subjuntivos de la forma «Si *estuviese*... entonces pensaría ...» son interpretados de forma diferente por diferentes filósofos —como lo son, por otra parte, todos los enunciados.

Un realista metafísico puede considerar que los enunciados que versan sobre *lo que sería el caso si*... son por sí mismos absolutamente verdaderos o falsos, independientemente de que en alguna ocasión *esté* justificado el aceptarlos o el rechazarlos. Si el relativista interpreta de este modo tan realista los enunciados que versan sobre lo que *creería* bajo tales-y-cuales condiciones, reconoce *una* clase de verdades absolutas, y deja de ser un relativista.

Un no-realista o un realista «interno» concibe los enunciados condicionales como enunciados que entendemos, en buena medida, aprehendiendo sus condiciones de *justificación* (como entendemos todos los demás enunciados). Esto no significa que el realista «interno» *abandone* la distinción entre verdad y justificación, sino que aprehendemos la verdad (justificación *idealizada*) tal y como aprehendemos cualquier otro concepto: por medio de una comprensión (en gran medida implícita) de los factores que hacen racionalmente aceptable decir que algo es verdadero. ¿Puede el *relativista* interpretar de esta forma no-realista, o realista «interna», los enunciados que se refieren a lo que *creería* en condiciones ideales?

Recordemos que la posición no-realista, tal y como la describí en el capítulo 3, asume una noción *objetiva* de aceptabilidad racional. El no-realista rechaza la opinión de que la verdad es una correspondencia con «un mundo prefabricado». Y es esto lo que le convierte en un *no*-realista (metafísico). Pero rechazar la teoría de la «verdad-

correspondencia» no es del todo lo mismo que considerar *subjetivas* tanto la verdad como la aceptabilidad. Nelson Goodman, quien concibe la verdad y la aceptabilidad racional como tipos de un predicado más general, «corrección», que se aplica tanto a las obras de arte como a los enunciados, ha expuesto el problema de un modo muy sucinto:

> Brevemente entonces, la verdad de los enunciados y la corrección de las descripciones, representaciones, ejemplificaciones, expresiones —de diseño, de dibujo, de lenguaje, de ritmo— es principalmente una cuestión de ajuste: ajuste a lo que nos referimos de diversas maneras, o ajuste a otras versiones, o ajuste a modos y maneras de organización. Las diferencias entre ajustar una versión a un mundo, ajustar un mundo a una versión y ajustar una versión a otra u otras versiones, se van desvaneciendo a medida que se reconoce el papel de las versiones en la elaboración de los mundos. Y se considera que el saber y el entendimiento se dan más allá de la adquisición de creencias verdaderas, hasta el descubrimiento y la ideación de ajustes de todo tipo.

Todo el *propósito* del relativismo, su propia característica definitoria, es, sin embargo, *negar* la existencia de cualquier noción inteligible de «ajuste» *objetivo.* Así que el relativista no puede entender la teoría de la verdad en términos de condiciones de justificación *objetivas.*

El intento de usar *condicionales* para explicar la distinción entre *estar en lo cierto y creer que se está en lo cierto* fracasa porque el relativista no tiene una noción *objetiva* de corrección para estos condicionales, como no la tiene para ningún otro tipo de enunciados.

Finalmente, si el relativista de hoy, como el viejo Protágoras, decidiese coger el toro por los cuerpos y afirmar que no hay diferencias entre «Estoy en lo cierto» y «Creo que estoy en lo cierto» —y afirmar que la distinción entre estar justificado y creer que se está justificado no puede ser trazada ni siquiera con respecto a *uno mismo*—, ¿qué significaría *hablar,* según esta concepción (más allá de producir ruidos con la esperanza de tener el *sentimiento subjetivo* de estar en lo cierto)? ¿En qué consiste pensar, más allá de producir imágenes-de-oración en la mente con la esperanza de tener el *sentimiento* subjetivo de estar en lo cierto? El relativista debe acabar negando que los pensamientos se *refieran* a algo, en el sentido realista o en el no-realista. En resumen, el relativista no se da cuenta de que la existencia de algún tipo de «corrección» objetiva es una presuposición del propio pensamiento.

Existe una interesante relación entre el argumento que acabo de esbozar (Platón-Wittgenstein) y el argumento contra la inconmensurabilidad que atribuía a Quine y Davidson: en efecto, ambos sostienen que un relativista consistente no debería tratar a los demás como hablantes (o como seres pensantes) —si sus «ruidos» son tan inconmensurables, entonces sólo son eso, ruidos—, mientras que Platón

y Wittgenstein sostienen que un relativista consistente no debería tratarse a sí mismo como un hablante o un sujeto pensante.

¿COMO SE EXPLICA ESTO?

Los argumentos que acabo de poner ante el lector me convencieron de que las dos filosofías de la ciencia más difundidas en este siglo son incoherentes. (Por supuesto, ninguna de ellas es *meramente* una «filosofía de la ciencia».) Naturalmente, esto me llevó a reflexionar sobre el significado de esta situación. ¿Cómo surgieron estos puntos de vista?

Recordaba antes que el positivismo lógico era tanto una continuación como algo distinto del positivismo machiano que le precedió. El positivismo de Mach, o «empirio-criticismo», era en realidad una repetición del empirismo humeano, expuesto en una jerga diferente. La brillantez de Mach, su estilo dogmático y entusiasta y su eminencia científica hicieron de su positivismo una difundida cuestión cultural (Lenin, temeroso de que los bolcheviques se convirtiesen al «empiriocriticismo», escribió un libelo contra éste). Einstein, cuya interpretación de la relatividad especial fue operacionalista en espíritu (en marcado contraste con la interpretación que ofreció de la relatividad general), reconocía que su crítica a la simultaneidad debía mucho a Hume y a Mach, a pesar de que, para su decepción, Mach rechazó totalmente la relatividad especial.

Pero el acontecimiento más notable de los que contribuyeron a la aparición del positivismo lógico fue la revolución en la lógica deductiva. Hacia 1879 Frege había descubierto un algoritmo, un procedimiento mecánico de prueba, que abarcaba hasta lo que hoy constituye la «lógica estándar de segundo orden». El procedimiento es *completo* para la teoría elemental de la deducción («lógica de primer orden»). El hecho de que se pueda apuntar un algoritmo para probar todas las fórmulas válidas de la lógica de primer orden —un algoritmo que no requiere un análisis semántico, ni tampoco una simulación de toda la psicología humana— es un hecho notable, que inspiró la esperanza de que se podía hacer lo mismo para la denominada «lógica inductiva» —«el método científico» podría acabar siendo un algoritmo— y que esos dos algoritmos —el algoritmo para la lógica deductiva (el cual resultó *incompleto* cuando se extendió a lógicas de orden superior) y el algoritmo a descubrir para la lógica inductiva— podrían describir exhaustivamente o «reconstruir racionalmente», no sólo la racionalidad *científica,* sino toda racionalidad digna de tal nombre.

Cuando estaba a punto de comenzar mi carrera docente en la Universidad de Princeton, conseguí conocer a Rudolf Carnap, que esta-

ba pasando dos años en el *Institute for Advanced Studies*. Una tarde memorable, Carnap me contó cómo había llegado a convertirse en un filósofo. Me explicó que había sido un licenciado en Física que estudiaba en el seminario de Frege. El texto era los *Principia Mathematica* (¡Imagínese, estudiar los *Principia* de Russell y Whitehead con Frege!). Carnap estaba fascinado tanto por la lógica simbólica como por la teoría de la relatividad especial. Así que decidió hacer su tesis sobre una formalización de la relatividad especial en la notación de los *Principia*. Carnap me confesó que llegó a convertirse en filósofo porque el departamento de Física de Jena no hubiera aceptado esta tesis.

Multitud de resultados negativos, incluyendo algunas poderosas generalizaciones debidas a Nelson Goodman, señalan hoy que *no puede* haber una lógica inductiva que sea completamente *formal*. Algunos aspectos importantes de la lógica inductiva pueden ser formalizados (aunque existe cierta controversia con respecto a la adecuación de la formalización), pero existe siempre la necesidad de juicios de «razonabilidad», estén éstos incorporados mediante la elección del vocabulario (o, con más precisión, mediante la *división* del vocabulario en predicados «proyectables» y predicados «no-proyectables») o de cualquier otra forma. Prácticamente nadie cree hoy en la existencia de un método científico puramente formal (v. Capítulo 8).

La historia que Carnap me contó apoya la idea de que el éxito de la formalización en el particular caso de la lógica deductiva desempeñó un papel crucial. Y si ese éxito inspiró el nacimiento del positivismo lógico, ¿no podría haber sido el fracaso en la formalización de la lógica inductiva, el descubrimiento de que no existe *algoritmo* alguno para la ciencia empírica, el que inspirara el nacimiento del «anarquismo»?

No insistiré en esta sugerencia; en cualquier caso, es probable que cooperasen factores adicionales. Mientras que Kuhn ha moderado su posición cada vez más, tanto Feyerabend como Michel Foucault han tendido a llevarlas hasta sus últimos extremos. Existe algo político en sus opiniones: tanto Feyerabend como Foucault vinculan nuestros criterios institucionalizados de racionalidad con el capitalismo, la explotación e incluso con la represión sexual. Sin duda son muchas las razones divergentes que hoy atraen a la gente hacia el relativismo extremado, y la idea de que todas las instituciones y tradiciones existentes hoy son inicuas es una de ellas.

Otra razón es cierto *cientifismo*. El carácter cientifista del positivismo es completamente abierto y sin tapujos, pero creo que también hay cierto cientifismo oculto bajo el relativismo. Pese a que los pensadores «anarquistas» nunca han abrazado por completo la teoría de que todo lo que cuenta para la «racionalidad» es aquello que establece nuestra cultura local, ésta constituye el límite natural de su tenden-

cia: y ésta es una teoría reduccionista. La teoría que afirma que la racionalidad se define por un programa computacional ideal es una teoría científica inspirada por las ciencias exactas; la que afirma que se define simplemente por normas culturales locales es una teoría científica inspirada por la antropología.

No voy a discutir aquí las esperanzas que los lingüistas chomskianos han despertado en algunos, las esperanzas en que la psicología cognitiva descubrirá algoritmos *innatos* que definan la racionalidad. Yo mismo pienso que es una moda intelectual que acabará decepcionándonos, tal y como nos decepcionó la esperanza del positivismo lógico con respecto a una lógica inductiva.

Todo ello sugiere que parte del problema en el que se halla la filosofía de nuestros días reside en cierto cientifismo heredado del siglo diecinueve —un problema que afecta a más de un campo intelectual. No niego que la lógica sea importante, ni que lo sean los estudios formales en teoría de la corroboración, en la semántica del lenguaje natural, etc. Me inclino a pensar que ocupan un lugar periférico con respecto a la filosofía, y que mientras nos paralice la formalización, es de esperar que continúe este tipo de movimiento pendular entre los dos tipos de cientifismo que he descrito. Ambos son intentos de eludir el problema de ofrecer una descripción equilibrada y humana del alcance de la razón.

6. HECHO Y VALOR

El asunto de los hechos y los valores, entendido de un modo suficientemente amplio, nos concierne a todos. En esto se diferencia netamente de muchos problemas filosóficos. La mayor parte de las mujeres y de los hombres cultos no se sienten en la obligación de adoptar una opinión definitiva con respecto al «problema» de si hay verdaderamente un mundo real o si sólo parece haberlo. Desde el punto de vista del modo de vida mayoritario, los problemas de la filosofía del lenguaje, de la epistemología y hasta de la metafísica pueden parecer asuntos que son interesantes, pero de elección opcional. Pero el problema de los hechos y los valores es de elección forzosa. Cualquier persona reflexiva *ha* de tener una opinión real con respecto a éste (opinión que puede coincidir o no con su opinión nocional). Y si la elección del problema de los hechos y valores es edundante forzosa, ocurre que una particular solución a este problema se ha arrogado el *status* de institución cultural, y es la siguiente: hechos y valores pertenecen a esferas totalmente distintas: la dicotomía «enunciados de hecho o enunciados de valor» es absoluta.

Al llamar a la dicotomía «institución cultural» pretendo sugerir que, por desgracia, la respuesta en cuestión seguirá siendo aceptada durante bastante tiempo, ignorando lo que los filósofos puedan decir sobre ella y sin tener en cuenta si es o no *correcta*. Aun cuando pudiera convencer al lector de que la dicotomía hecho-valor carece de base racional, o aun cuando algún filósofo mejor que yo pudiera demostrarlo con un argumento absolutamente conclusivo (en filosofía no existen tales argumentos, por supuesto), aun así, la próxima vez que el lector salga a la calle, o a tomar unas copas, o mantenga una discusión en un colectivo de debate, se encontrará alguien que le diga «¿He de suponer que es un enunciado de hecho o un juicio de valor?». La opinión según la cual no hay base objetiva para decidir si las cosas son buenas o malas, o mejores o peores, se ha convertido en algo *institucionalizado.*

La estrategia de mi argumento no es nueva. Voy a rehabilitar un argumento un tanto desacreditado en el debate sobre hechos y valores, a saber, el que defiende que esa distinción es al menos irremediablemente difusa, ya que los propios enunciados fácticos y los procedimientos de investigación empírica con los que contamos para decidir si algo es o no es un hecho presuponen valores.

El descrédito de este argumento se debe a una réplica obvia. La réplica al argumento que apela a la presuposición de valores por par-

te de la ciencia constituye una concesión preventiva: los defensores de la dicotomía hecho-valor conceden que la ciencia presupone valores, por ejemplo, presupone que queremos la *verdad,* pero sostienen que esos valores no son *éticos.* Imaginaré un oponente un tanto débil que adopte la opinión de que la ciencia presupone *un* valor, a saber, el propio valor de la verdad.

Como hemos visto, la verdad no es una noción simple. La idea conforme a la cual la verdad es una copia pasiva de lo que «realmente existe» (con independencia de la mente y del discurso) se ha derrumbado bajo las críticas de Kant, Wittgenstein y otros filósofos, si bien continúa ejerciendo una profunda influencia sobre nuestra forma de pensar.

Algunos filósofos han apelado al *principio de equivalencia* —esto es, el principio que reza: *decir que un enunciado es verdadero equivale a afirmar el enunciado*— para negar que existan problemas filosóficos con respecto a la verdad. Otros apelan a la obra de A. Tarski, el lógico que mostró de qué forma, dado un lenguaje formalizado (una notación formal en la cual expresar ciertos enunciados, empleando la lógica simbólica), podemos definir «verdadero» *para ese lenguaje* en un lenguaje de orden superior (denominado «metalenguaje»)[1].

La propia obra de Tarski está basada en el principio de equivalencia: de hecho, su criterio para obtener una definición adecuada de «verdad» consistía en que ésta había de ofrecer como teoremas del metalenguaje todas aquellas oraciones de la forma *«P» es verdadera si y sólo si P* (donde *P* es una oración de la notación formal en cuestión), por ejemplo,

(T) «La nieve es blanca» es verdadera si y sólo si la nieve es blanca.

No obstante, el principio de equivalencia es filosóficamente neutral, como lo es la obra de Tarski. Según *cualquier* teoría de la verdad, «La nieve es blanca» es equivalente a «"La nieve es blanca" es verdadera».

Los filósofos positivistas replicarían que si usted sabe *(T),* sabe lo que significa «"La nieve es blanca" es verdadera»: significa *La nieve es blanca.* Y añadirán que si usted no comprendiese «nieve» y «blanca», se hallaría en un serio apuro. Pero el problema no es que no comprendamos «La nieve es blanca»; el problema es que no comprendemos *qué es comprender* «La nieve es blanca». Este es el problema filosófico. Con respecto a este problema, *(T)* no dice nada.

Y, realidad, ¿no concuerda esto último con nuestras intuiciones sobre estas cuestiones? Si se nos aproxima alguien con un brillo especial en sus ojos y nos pregunta «¿Quieren ustedes saber la "Verdad"?», la reacción más normal es mirar a esta persona con cierta suspicacia.

---

[1] Para un estudio no-técnico de la obra de Tarski véase mi libro *Meaning and the Moral Sciences,* Parte I, sección I.

Y el motivo de nuestra suspicacia (aparte del brillo especial de sus ojos) es precisamente que cuando alguien nos está diciendo que quiere que sepamos la verdad, en realidad no nos dice *nada* en tanto que no tengamos ni idea de los estándares de aceptabilidad racional a los que esa persona se adhiere: qué constituye para él un modo racional de seguir una investigación, cuáles son sus estándares de objetividad, cuándo considera racional dar por acabada una investigación, qué consideraciones estima que proporcionan una buena razón para aceptar un veredicto u otro, sea cual sea el tipo de problema en el que pueda estar interesado. Aplicado al caso de la ciencia, yo diría que afirmar que ésta «busca descubrir la verdad» es en realidad establecer un enunciado puramente formal. Es decir tan sólo que los científicos no quieren aseverar que la nieve es blanca si la nieve no es blanca, ni que hay electrones fluyendo a través del cable si los electrones no están fluyendo a través del cable, etc. Pero estos enunciados puramente formales son completamente vacíos mientras no tengamos ni idea de cuál es el sistema de criterios de aceptabilidad racional que distingue las tentativas científicas de determinar que la nieve es blanca de las que no lo son, las tentativas científicas de determinar si los electrones están fluyendo a través del cable de las que no lo son, etc.

Si la idea de comparar nuestro sistema de creencias con una realidad no conceptualizada para ver si se «emparejan» no tiene sentido, entonces la afirmación: «la ciencia busca descubrir la verdad» sólo puede significar que la ciencia constituye una imagen del mundo que, en el límite ideal, satisface ciertos criterios de aceptabilidad racional. Que la ciencia busca construir una imagen del mundo que sea *verdadera* es en sí mismo un enunciado verdadero, un enunciado verdadero, formal y casi vacío; sólo los criterios de aceptabilidad racional implícitos en la ciencia dotan de contenido material a sus objetivos. En resumen, estoy diciendo que la respuesta a la posición del oponente imaginario, aquélla que defiende que el único objetivo de la ciencia es descubrir la verdad (además de señalar que la ciencia tiene objetivos adicionales, lo cual es cierto, claro está) es que la verdad no es la *cuestión de fondo:* la verdad misma obtiene su vida a partir de nuestros criterios de aceptabilidad racional, y debemos examinar éstos si deseamos descubrir los valores que están efectivamente implícitos en la ciencia.

Permítaseme imaginar, a modo de ejemplo, un caso extremo de desacuerdo. No voy a imaginar un desacuerdo científico ordinario, aunque espero que nuestra respuesta nos permita descubrir algo acerca de la naturaleza de los valores científicos.

En el caso que estoy a punto de describir el desacuerdo tendrá que ver con la hipótesis que discutimos en el capítulo 1, la hipótesis según la cual todos somos cerebros en una cubeta. Hemos argumentado que no es posible que esta hipótesis sea verdadera, pero supondremos que

nuestros argumentos han fracasado a la hora de convencer a una de las partes en disputa (lo cual no es improbable, ya que los argumentos filosóficos nunca convencen a todos). En resumen, la hipótesis afirma que todo es una alucinación, tal y como describimos anteriormente.

Aunque todo fuese una alucinación colectiva, mucha gente no tendría por qué notarlo. Para los amantes casi no constituiría ninguna diferencia, por ejemplo[2]. E imaginemos que tampoco constituiría ninguna diferencia para los economistas. (¿Por qué un economista tendría que sentirse preocupado por el hecho de que todo el dinero no fuese físicamente real? La mayor parte de éste no es físicamente real en ninguna teoría.)

Quiero que el lector se imagine que esta disparatada (y yo diría que incoherente) teoría no la mantiene un lunático aislado, sino todos los habitantes de un gran país, Australia, por ejemplo. Imaginemos que sólo una pequeña minoría de los habitantes de Australia tienen creencias parecidas a las nuestras, mientras que la gran mayoría cree que somos cerebros en una cubeta. Y quizás los australianos tienen esta creencia debido a que son discípulos de un gurú, el gurú de Sidney. Tal vez, si les expresásemos nuestras dudas, nos responderían: «Oh, si ustedes pudiesen hablar con el gurú de Sidney y ver lo bueno, amable y sabio que es, también se convencerían». Y si les preguntásemos: «Pero, si la ilusión es tan perfecta como usted dice, ¿cómo sabe el gurú de Sidney que somos cerebros en una cubeta?», podrían contestar: «Oh, el gurú de Sidney simplemente *lo sabe*».

Tal y como he dicho, éste no es un desacuerdo científico en el sentido ordinario. Podemos imaginar que los australianos son tan buenos como nosotros a la hora de anticipar experiencias, de construir puentes que se sostienen (o que parecen sostenerse), etc. Hasta pueden estar deseando aceptar nuestros últimos descubrimientos científicos, no como verdaderos, sino como descripciones correctas de lo que parece pasar en la imagen. Podemos o no imaginar que discrepan con nosotros con respecto a algunas predicciones concernientes al futuro lejano (por ejemplo, pueden esperar que algún día se averíe la maquinaria automática y entonces la gente comience a tener alucinaciones de la misma clase, que acabarían proporcionándoles evidencia de que su concepción es la correcta)[3], pero el que lleven a cabo tales pre-

[2] Pero me reservo la posibilidad de cambiar de opinión con respecto a si constituiría o no una diferencia.

[3] El hecho de que realizasen tales predicciones marcaría esta diferencia: su concepción ya no sería incoherente tal y como criticamos en el Capítulo I, ya que estarían realizando afirmaciones que podrían ser justificadas (eventualmente) y, por tanto, su concepción no requeriría una teoría de la verdad «trascendente» (o independiente de la justificación) para ser comprendida.

dicciones o que se comprometan a no realizar predicciones distintas de las que les permite la teoría estándar no afectará a mi argumento. La cuestión es que hemos concebido un caso en el que un vasto número de personas tiene un sistema de creencias que es completo en sí mismo, pero que discrepa violentamente del nuestro.

En este punto no hay ningún problema de desacuerdo con respecto a los valores «éticos»; los australianos pueden tener una ética tan similar a la nuestra como a usted se le antoje. (Aunque un antiguo griego hubiese dicho que ser *sabio* es un valor *ético;* de hecho, el judaísmo y el cristianismo han estrechado el concepto de lo ético debido a cierta concepción de la salvación.)

Lo primero que quiero observar con respecto a estos hipotéticos autralianos es que su visión del mundo es «*una locura*». Sin duda, este término se usa en algunas ocasiones casi como un término de aprobación; pero en este punto no le atribuyo ese sentido. Creo que observaríamos con gran tristeza a una comunidad de seres humanos que abrazase una visión del mundo tan insana. Consideramos dementes a los australianos en el sentido de que tienen mentes *enfermas,* y esta caracterización de sus mentes es ética, o raya en lo ético. Pero, aparte de atribuyéndoles calificativos, ¿existe otra forma de discutir con los australianos? (O de intentar discutir, puesto que supondré que no pueden ser convencidos.)

Podemos concebir de inmediato un argumento que tiene que ver con la *incoherencia* de su concepción. No me refiero a la incoherencia de la que hablamos en el capítulo 1. Aquella era una incoherencia de *fondo,* cuya exposición requería un argumento filosófico (y, por ende, discutible). La concepción australiana es incoherente a nivel mucho más superficial. Uno de nuestros objetivos es ser capaces de ofrecer una descripción de cómo sabemos que nuestros enunciados son verdaderos. Lo intentamos, en parte, avanzando una teoría causal de la percepción, de modo que podemos describir lo que nosotros tomamos por fiabilidad de nuestro conocimiento perceptivo, considerado desde dentro de nuestra propia teoría, dando una descripción, dentro de la teoría, de cómo las percepciones resultan de la actuación de nuestros órganos transductores sobre el mundo externo. Y, en parte, mediante una teoría de estadística y una estrategia experimental, de forma que podemos demostrar, dentro de nuestra propia teoría, cómo los procedimientos que adoptamos para excluir el error experimental tienden a excluirlo en la mayoría de los casos. En resumen, es un constreñimiento importante y sumamente útil el que nuestra teoría en evolución del mundo, considerada como un todo, incluya una descripción de las actividades y procesos mediante los que podemos saber que es correcta.

No obstante, el sistema de los australianos no posee esta propiedad de coherencia (al menos tal como nosotros la consideramos, y

la «coherencia» no es algo para lo que tengamos un algoritmo, sino algo que en última instancia juzgamos «a ojo de buen cubero»). Recordemos que los australianos postulan una ilusión tan perfecta que no existe ningún modo racional gracias al cual el gurú de Sidney pueda *saber* que el sistema de creencias que él ha adoptado (y ha instado a adoptar a los demás) es correcto. Juzgado a partir de nuestros estándares de coherencia, *su* sistema de creencias es totalmente incoherente.

Pueden apuntarse otras virtudes metodológicas de las que carece su sistema de creencias. Este último, tal y como lo he descrito, concuerda con el nuestro en las leyes de la naturaleza, tal como aparecen *en la imagen*. Pero ¿nos dice si las leyes que parecen cumplirse en la imagen son las leyes de la naturaleza que efectivamente se cumplen fuera de la cubeta? Si no lo hace, entonces carece de cierto tipo de alcance comprensivo que perseguimos, pues no nos dice, ni siquiera en sus propios términos, cuáles son las verdaderas leyes últimas de la naturaleza. Viola la navaja de Ockham, desde luego. Una vez más, parece difícil o imposible formalizar la navaja de Ockham en un algoritmo, pero el mismo hecho de que la teoría de los cerebros en una cubeta postule toda suerte de objetos fuera de la cubeta, sin que desempeñen papel alguno en la explicación de sus experiencias, de acuerdo con su propia teoría, deja claro que podemos afirmar categóricamente que en este caso se viola la máxima... «no multiplicar las entidades sin necesidad». Llamaremos *funcionalmente simple* a una teoría que obedezca a la navaja de Ockham, más en el espíritu que en la letra.

Lo que he venido afirmando es que los procedimientos mediante los que decidimos la aceptabilidad de una teoría científica tienen que ver con que la teoría científica, considerada como un todo, exhiba ciertas «virtudes» o no las exhiba. Estoy suponiendo que el procedimiento de construcción de una teoría científica no puede analizarse correctamente como un procedimiento de verificación de teorías científicas *oración por oración*. Estoy suponiendo que la verificación científica es una cuestión holística, que son los sistemas científicos enteros los que se enfrentan al tribunal de la experiencia «como un cuerpo integrado», y que el juicio resultante es una cuestión un tanto intuitiva, que no puede formalizarse a menos que formalicemos toda la psicología humana. Pero volvamos a nuestra pregunta inicial. ¿Cuáles son los valores implícitos de la ciencia?

Mi argumento establece que si consideramos los valores a los que apelamos en nuestra crítica a los cerebro-cubetistas (y añadimos, desde luego, otros valores que no se discuten en este caso, por ejemplo, nuestro deseo de eficacia instrumental, presumiblemente compartido con los cerebro-cubetistas) entonces obtenemos una imagen de la ciencia en la que ésta presupone un rico sistema de valores. El hecho es que, si consideramos el ideal de aceptabilidad racional que se revela

al examinar cuáles son las teorías científicas que tanto los científicos como la gente de la calle consideran racionalmente aceptables, entonces observaremos que el propósito de nuestra actividad científica es construir una representación del mundo que tenga las siguientes características: eficacia instrumental, coherencia, alcance comprensivo y simplicidad funcional. Pero ¿por qué razón?

Yo respondería que la razón por la cual queremos este tipo de representación, y no el «tipo enfermo» de mundo nocional que poseen los australianos y los cerebro-cubetistas, es que este tipo de representación forma parte de nuestra idea de florecimiento cognitivo humano y, por tanto, parte de nuestra idea de florecimiento humano total o *Eudaemonia.*

Desde luego, si el realismo metafísico fuera correcto, si fuese posible contemplar el objetivo de la ciencia como un intento de «emparejar» nuestro mundo nocional con el mundo en-sí-mismo, entonces se podría afirmar que sólo nos interesa la coherencia, el alcance comprensivo, la simplicidad funcional y la eficacia instrumental en tanto que instrumentos que sirven al fin de lograr este «emparejamiento». Pero el concepto de «emparejamiento transcendental» entre nuestras representaciones y el mundo en sí mismo no tiene sentido. Con todo, negar que busquemos este tipo de igualación metafísica con un mundo metafísico no es negar que busquemos el tipo usual de ajuste empírico (decidida por nuestros criterios de aceptabilidad racional) con un mundo *empírico.* Pero el mundo empírico, en oposición al mundo nouménico, depende de nuestros criterios de aceptabilidad racional (y viceversa, desde luego). Utilizamos nuestros criterios de aceptabilidad racional para elaborar una imagen teórica del «mundo empírico» y conforme se desarrolla esta imagen revisamos bajo su luz nuestros propios criterios de aceptabilidad racional, y así sucesiva e ininterrumpidamente. En mis otros libros he puesto énfasis en la dependencia de nuestros métodos con respecto a nuestra imagen del mundo. Lo que quiero resaltar aquí es la otra cara de la moneda, la dependencia del mundo empírico con respecto a nuestros criterios de aceptabilidad racional. Lo que trato de afirmar es que hasta para tener un mundo empírico debemos tener criterios de racionalidad, y que éstos revelan parte de nuestra concepción de una inteligencia especulativa óptima. En resumen, estoy afirmando que el «mundo real» depende de nuestros valores (y, una vez más, también al contrario).

AL MENOS ALGUNOS VALORES DEBEN SER OBJETIVOS

El hecho de que la ciencia no sea, como se ha pensado, «neutral en cuanto a los valores», no muestra ni que los valores «éticos» sean

objetivos, ni que la ética pueda ser una ciencia, claro está. De hecho, no existe ninguna probabilidad de alcanzar una «ciencia» de la ética, sea en el sentido de una ciencia de laboratorio o en el sentido de una ciencia deductiva. Como Aristóteles señaló hace tiempo[4],

> Por consiguiente, hablando de cosas de esta índole y con tales puntos de partida, hemos de darnos por contentos con mostrar la verdad de un modo tosco y esquemático; hablando sólo de lo que ocurre por lo general y partiendo de tales datos, basta con llegar a conclusiones semejantes. Del mismo modo se ha de aceptar cuanto aquí digamos: porque es propio del hombre instruido buscar la exactitud en cada género de conocimientos en la medida en que la admite la naturaleza del asunto; evidentemente, tan absurdo sería aprobar a un matemático que empleara la persuasión como reclamar demostraciones a un retórico.

Pero el hecho de que la aceptabilidad racional en las ciencias exactas (que son sin duda los ejemplos centrales del pensamiento racional) dependa de *virtudes cognitivas* como la «coherencia» y la «simplicidad funcional» muestra que al menos algunos términos de valor representan propiedades de las cosas a las que se aplican, y no precisamente sentimientos de la persona que los usa.

Si los términos «coherente» y «simple» no están por *propiedades* de las teorías, ni siquiera por propiedades difusa o imperfectamente definidas, sino sólo por «actitudes» que ciertas personas mantienen hacia esas teorías, entonces los términos que guardan relación con la aceptabilidad racional, tales como «justificado», «bien confirmado» o «la mejor explicación de que disponemos», deben ser subjetivos casi por completo: pues la aceptabilidad racional no puede ser más objetiva que los parámetros de los que depende. No obstante, como ya indicamos en el capítulo precedente, la concepción que mantiene que la aceptabilidad racional es sencillamente subjetiva se refuta a sí misma. De modo que nos vemos obligados a concluir que al menos estos términos de valor tienen algún tipo de aplicación objetiva, esto es, algún tipo de condiciones objetivas de justificación.

Por supuesto, uno podría intentar zafarse de admitir la existencia de valores objetivos de cualquier clase negando que «coherente», «simple», «justificado», y otros términos, sean términos de *valor*. Uno podría mantener que estos términos representan propiedades que nosotros valoramos, pero que no hay corrección objetiva en relación con esta práctica. No obstante, esta tentativa tropieza con dificultades de inmediato. «Coherente» y «simple» comparten demasiadas características con los términos de valor paradigmáticos: «coherente» y «simple», como «amable», «bello» y «bueno», se emplean a menudo co-

---

[4] *Ethica Nicomachea,* libro I, cap. 3, trad. cast. M. Araujo y J. Marías, Instituto de Estudios Políticos, Madrid, 1970.

mo términos de *elogio*. Nuestras concepciones de la coherencia, de la simplicidad y de la justificación están condicionadas históricamente en la misma medida en que lo están nuestras concepciones de la amabilidad, la belleza y la bondad; estos términos epistémicos figuran en los mismos tipos de perenne controversia filosófica en los que figuran los términos que representan valores éticos y estéticos. Es obvio que la concepción de la racionalidad del cardenal John Newman era completamente distinta de la de Rudolf Carnap. Aun cuando hubiesen vivido en la misma época y hubiesen estado dispuestos al diálogo, es altamente improbable que uno pudiera haber convencido al otro. La pregunta *¿cuál es la concepción racional de la racionalidad?* es difícil en la misma medida en que es difícil justificar un sistema ético. No hay ninguna concepción *neutral* de la racionalidad a la que apelar.

En este punto podrían ensayarse varias maniobras convencionalistas, por ejemplo, considerar que «$\text{justificado}_{\text{CARNAP}}$» y «$\text{justificado}_{\text{NEWMAN}}$» son *propiedades* diferentes, y que la decisión de referirse con la palabra «justificado» a «$\text{justificado}_{\text{CARNAP}}$» o a «$\text{justificado}_{\text{NEWMAN}}$» implica un «juicio subjetivo de valor», pero que consignar el hecho de que un enunciado dado *S* está $\text{justificado}_{\text{CARNAP}}$ o $\text{justificado}_{\text{NEWMAN}}$ no implica juicio de valor alguno. ¿Pero desde qué punto de vista se está empleando la palabra *«hecho»?* Si no hay una concepción de la racionalidad que *debamos* tener objetivamente, la noción de «hecho» es vacía. Sin los valores cognitivos de coherencia, simplicidad y eficacia instrumental, carecemos de mundo y de «hechos», hasta de hechos con respecto a aquellas cosas que son *relativas* y a qué lo son, porque éstos se hallan en el mismo barco que todos los demás hechos. Pero estos valores cognitivos son arbitrarios salvo que se consideren parte de una concepción holística del florecimiento humano. Una vez privados de la vieja idea realista de la «verdad como correspondencia» y de la idea positivista bajo cuya óptica la justificación se fija mediante «criterios» públicos, nos hemos quedado con la necesidad de considerar nuestra búsqueda de mejores concepciones de la racionalidad como una actividad intencional y humana, la cual, como cualquier actividad que se alce por encima del hábito y del mero seguimiento de la inclinación o de la obsesión, está orientada por la idea de lo bueno.

LA RACIONALIDAD EN OTRAS AREAS

Si los valores implícitos en la ciencia, y especialmente en las ciencias exactas, revelan parte de nuestra idea de lo bueno, creo que la parte restante puede descifrarse a partir de nuestros estándares de aceptabilidad racional en otras áreas de conocimiento. Aunque en este pun-

to es necesario ampliar la noción de estándares *de aceptabilidad racional,* claro está.

Hasta ahora sólo hemos examinado los estándares de aceptabilidad racional en un sentido literal; estándares que nos dicen cuándo hemos de aceptar enunciados y cuándo no hemos de aceptarlos. Ahora bien, en un sentido más amplio, los estándares de aceptabilidad racional tienen que ver con el modo en que juzgamos no sólo la verdad y la falsedad de los sistemas de enunciados, sino también su *adecuación y perspicuidad.* Hay otros motivos —estrictamente cognitivos— por los que un sistema de enunciados puede ofrecernos una descripción que no resulte suficientemente satisfactoria, además de por ser falso.

De haberlo querido, podría haber puesto de relieve este punto en conexión con la ciencia teorética. Podría haber señalado que el interés de la ciencia exacta no es tan sólo descubrir enunciados verdaderos, incluso enunciados verdaderos y universales («Leyes»), sino encontrar enunciados que sean verdaderos y *relevantes.* Y el concepto de *relevancia* arrastra tras de sí un amplio grupo de intereses y valores. Con todo, ello sólo habría sido apuntar que nuestro *conocimiento* del mundo presupone valores, y no establecer la afirmación más radical de que lo que *cuenta* como mundo real depende de nuestros valores.

Cuando llegamos a la racionalidad perceptiva, esto es, a las técnicas y criterios implícitos sobre cuya base decidimos si alguien puede ofrecernos un informe verdadero, adecuado y perspicaz hasta de los hechos *perceptivos* más simples, vemos que entran en juego un gran número de factores. Recientemente, los psicólogos han hecho hincapié en que hasta los casos más simples de percepción envuelven bastante construcción teórica. Esto no sólo es cierto a nivel neurofisiológico, sino también a nivel cultural. Una persona perteneciente a una cultura que carece de muebles podría ofrecer cierto tipo de descripción de una habitación, pero si no sabe lo que es una mesa, una silla o un pupitre, su descripción difícilmente comunicaría la información sobre la habitación que un miembro de nuestra cultura esperaría obtener. Su descripción podría constar de enunciados verdaderos, y aún así, no sería la descripción adecuada. Este simple ejemplo muestra ya que el requisito de que una descripción sea adecuada equivale de forma implícita al requisito de que el descriptor pueda disponer de cierto conjunto de conceptos: esperamos que los descriptores racionales sean capaces de adquirir ciertos tipos de conceptos con respecto a ciertos tipos de descripción, y que comprendan que su uso es necesario en esos casos; el hecho de que el descriptor no emplee cierto concepto puede ser motivo para que tanto él como su descripción sean criticados.

Lo que es cierto en un nivel tan simple como el de hablar acerca

de las mesas y las sillas de una habitación deshabitada, lo es también en el nivel de la descripción de las relaciones y situaciones interpersonales. Examinemos los términos que usamos todos los días para describir lo que nos parecen las demás personas, por ejemplo, *amables o poco amables*. Estos términos pueden usarse, cómo no, para elogiar o censurar; y una de las muchas distinciones que se han confundido bajo el rótulo general «distinción hecho-valor» es la distinción entre usar una expresión lingüística para describir y usarla para elogiar o censurar. Mas esta distinción no puede trazarse sobre la base del vocabulario. El juicio que afirma que alguien es poco amable puede emplearse para censurar; pero puede emplearse sencillamente para describir, y también para explicar o predecir.

Por ejemplo, podría advertirle al lector: «No deje que Hernández hiera sus sentimientos. Su forma de hablar puede insinuar cierta antipatía, pero esta impresión errónea es bastante normal. Es probable que se siga comportando de un modo que hiera sus sentimientos, sienta lo que sienta por usted. Es un hombre poco amable, pero no piense que tiene algo contra usted».

En este pequeño e imaginario discurso, estoy empleando las palabras «poco amable» sin el propósito de censurar a Hernández, sino con la intención de predecir y explicar a alguien su comportamiento. Y tanto la predicción como la explicación pueden ser perfectamente *correctas*. De forma similar «celoso» puede ser un término de censura, pero también puede usarse sin intención de censurar. (Algunas veces uno tiene perfecto derecho a estar celoso.)

El uso de la palabra «amable» me parece un buen ejemplo de cómo la distinción hecho-valor es desesperadamente difusa en el mundo y en el lenguaje reales. En *The Sovereignty of «Good»* [5] Iris Murdoch ha puesto de manifiesto la importancia que tienen términos como «poco amable», «impertinente», «obstinado», «molesto», etc., en la evaluación moral efectiva. Aun cuando cada uno de los siguientes enunciados, «Juan es un hombre muy poco amable», «Juan es un hombre que sólo piensa en sí mismo», «Juan haría prácticamente cualquier cosa por dinero», puede ser una descripción verdadera en el sentido más positivista (y démonos cuenta de que «Juan haría prácticamente cualquier cosa por dinero» no contiene ningún término de valor) si uno establece la conjunción de los tres enunciados, no es necesario añadir que «Juan no es muy buena persona». Cuando concebimos hechos y valores como cosas independientes, lo típico es pensar que los «hechos» se enuncian en alguna jerga fisicalista o burocrática, y que los «valores» necesitan de términos de valor más abstractos, por ejemplo, «bueno», «malo». Pero es difícil seguir mantenien-

[5] Routledge and Kegan Paul, 1970.

do la independencia de los hechos con respecto a los valores cuando los mismos hechos son del tipo de «poco amable», «piensa solamente en sí mismo», «haría cualquier cosa por dinero». Así como criticamos a un descriptor que no emplee los conceptos de *mesa* y *silla* cuando se le exige su uso, quien no observe que alguien es *amable* o *espontáneo* puede exponerse a ser tachado de imperceptivo o superficial; su descripción no es una descripción adecuada.

LOS SUPERBENTHAMITAS

Permítaseme retroceder y modificar mi anterior ejemplo de los «cerebro-cubetistas». Imaginemos esta vez que el continente australiano está poblado por una cultura cuya historia, geografía y ciencia exacta son semejantes a las nuestras, pero cuya ética discrepa de la que nosotros mantenemos. No quiero servirme del habitual ejemplo de los supernazis, sino considerar, mas bien, el caso de los superbenthamitas, de mayor interés. Imagine que el continente australiano está poblado por gentes que tienen una medida científica y detallada de lo que ellos suponen que es el «tono hedónico», y que creen que uno ha de obrar siempre con vistas a maximizar el tono hedónico (considerando que ello significa conseguir el mayor tono hedónico para el mayor número de gente). Supondré que los superbenthamitas son extremadamente sofisticados, y son conscientes de todas las dificultades que entraña la predicción del futuro, la estimación exacta de las consecuencias de las acciones, etc. Supondré también que son extremadamente despiadados, y aunque no harían sufrir a nadie para proporcionar la mayor felicidad al mayor número de gente si hubiesen dudas razonables de que sus acciones fuesen a tener realmente como consecuencia producir la mayor felicidad al mayor número de gente, en aquellos casos en los que sí puede saberse con certeza cuáles serían las consecuencias de las acciones, los superbenthamitas estarían dispuestos a llevar a cabo las acciones más atroces —estarían dispuestos a torturar niños o a condenar a personas por crímenes no cometidos— si el resultado de esas acciones fuese incrementar a la larga el nivel general de satisfacción (después de administrar a la víctima inocente la ración de sufrimiento oportuna en cada caso) en cualquier número positivo E, por muy pequeño que sea.

Me imagino que no nos sentiríamos muy satisfechos con esta especie de moralidad superbenthamita. La mayoría condenaríamos a los superbenthamitas por su sistema de valores enfermo, por burócratas, por despiadados, etc. Son el «hombre nuevo» en su manifestación más nefasta. Y ellos nos devolverían nuestras invectivas diciéndonos que somos unos bobos supersticiosos, que somos prisioneros de una tradición irracional, etc.

El desacuerdo existente entre nosotros y los superbenthamitas es precisamente el tipo de desacuerdo que generalmente se imagina en orden a señalar que dos grupos de personas podrían estar de acuerdo acerca de la totalidad de los hechos y, pese a todo, no estarlo con respecto a los «valores». Pero examinemos el caso más de cerca. Cada superbenthamita está familiarizado con el hecho de que la mayor satisfacción para el mayor número de personas (medida en «útiles») exige a veces decir una mentira. Y decir mentiras para maximizar el nivel general de placer no se considera *deshonesto,* en sentido peyorativo. Así que, al poco tiempo, el uso de la descripción «honesto» entre los superbenthamitas será en extremo diferente de nuestro uso del mismo término descriptivo. Y lo mismo pasará con «amable», «buen ciudadano», etc. Con el tiempo, el vocabulario del que disponen los superbenthamitas para la descripción de las situaciones interpersonales será completamente distinto del nuestro. Y no sólo carecerán de nuestros recursos descriptivos, o los habrán alterado hasta hacerlos irreconocibles, sino que es posible que inventen su nueva jerga (por ejemplo, términos exactos para describir los tonos hedónicos) a la que no tendremos acceso. La textura del mundo humano comenzará a cambiar. Con el curso del tiempo, los superbenthamitas acabarán viviendo en un mundo humano diferente del nuestro.

En resumen, no es cierto que «coincidamos con los superbenthamitas en los hechos y discrepemos con respecto a los valores». En relación con casi todas las situaciones interpersonales, nuestra descripción de los hechos será completamente distinta de la suya. Aun cuando ninguno de sus enunciados acerca de la situación fuera *falso,* no consideraríamos a su descripción ni adecuada ni perspicua. Y ellos imputarían los mismos defectos a la nuestra. En resumen, aun cuando dejemos a un lado nuestro «desacuerdo con respecto a los valores» no podríamos concebir su representación del mundo como una representación completa y racionalmente aceptable. Y así como la incapacidad de los cerebro-cubetistas para comprender correctamente *cómo es el mundo* es resultado directo de la enfermedad de sus estándares de racionalidad (teorética), la incapacidad de los superbenthamitas para comprender correctamente cómo es el mundo humano es un resultado directo de *su* concepción enferma del florecimiento humano.

SUBJETIVISMO CON RESPECTO A LA BONDAD

Se ha afirmado con frecuencia que el paso de «Juan es veraz, amable, valeroso, responsable, etc.» a «Juan es moralmente bueno» implica al menos una «premisa» no demostrada (e indemostrable), a saber, «La amabilidad es moralmente buena». Y se ha sostenido que

la necesidad de «premisas morales» para extraer conclusiones morales a partir de enunciados «fácticos» muestra que los enunciados éticos no son racionalmente justificables. Esta imagen de la ética como una especie de pirámide invertida, con un vértice (que carece de apoyos) consistente en «axiomas éticos» y que cimenta todo nuestro cuerpo de creencias y pensamientos morales, es ingenua. Nadie ha logrado imponer una estructura axiomática sobre la ética (como Aristóteles observaba en el pasaje citado pocas páginas atrás, las máximas morales que podamos apuntar serán casi siempre verdaderas tan sólo «por lo general»). Los escépticos han empleado la misma estrategia en todas las áreas: describir un cuerpo de conocimientos sobre el que se desea arrojar dudas de modo que descanse sobre «axiomas» insostenibles. Los escépticos que dudan de la existencia de objetos materiales, por ejemplo, sostienen que el principio: «Si nuestras sensaciones se producen como si hubiera un mundo material, entonces probablemente exista un mundo material», es una *premisa racionalmente insostenible* a la que recurrimos tácitamente cuando decimos «observar» un objeto material, o, por otra parte, cuando intentamos justificar la creencia en su existencia. En realidad, la ética, la matemática y el discurso sobre objetos materiales presuponen conceptos, no «axiomas». Los conceptos se usan en la observación y en la generalización, y están legitimados por el éxito que logramos al utilizarlos en esas actividades.

Un ataque más sofisticado a la idea de objetividad ética admite que nuestras creencias éticas descansan en observaciones de casos específicos, «intuiciones», máximas generales, etc., y no en un cúmulo de «axiomas éticos» arbitrarios, pero formula la acusación que establece que las propias observaciones éticas están afectadas por una incurable dolencia: la *proyección.*

De acuerdo con esta descripción, los humanos somos por naturaleza compasivos, si bien no siempre. De este modo, cuando observamos algún horrible suceso, como podría ser que alguien estuviese torturando a un niño pequeño únicamente por sadismo, nos horrorizamos (a veces). Pero el mecanismo psicológico de la «proyección» nos lleva a experimentar la cualidad sentida como una cualidad de la propia acción: decimos «Fue un acto horroroso», cuando realmente deberíamos decir: «Mi reacción fue sentirme horrorizado». De esta manera vamos construyendo un cuerpo de lo que nosotros consideramos observaciones éticas, pero que son en realidad tan sólo observaciones con respecto a nuestros propios *sentimientos* éticos subjetivos.

Este relato tiene formas más sofisticadas (como cualquier otro). Hume postuló una tendencia humana a la que llamó «simpatía», tendencia que ha ido ampliándose gradualmente bajo la influencia de la cultura. Los sociobiólogos contemporáneos postulan un instinto al que llaman «altruísmo» y hablan de «genes altruístas». Pero la idea clave

sigue siendo la misma: hay *sentimientos* éticos, pero no propiedades objetivas de valor.

Ya hemos visto que esto no es cierto: hay al menos *algunos* valores objetivos, por ejemplo, la *justificación*. Con todo, podría afirmarse que los valores *éticos* son subjetivos, mientras que los valores *cognitivos* son objetivos; pero el argumento que establece que es absolutamente imposible que hayan valores objetivos ha sido refutado.

Para mostrar en qué se equivocan los argumentos en favor del subjetivismo moral, debo traer a la memoria los argumentos que usábamos en el capítulo 2 en contra del realismo metafísico. Esto puede parecer extraño: ¿no es el subjetivismo lo *contrario* del realismo metafísico? Si alguien cree que es así, también creerá que cualquier argumento contra el realismo metafísico debe *respaldar* al subjetivismo; la estrategia que voy a seguir, a saber, emplear el *mismo* argumento tanto contra el realismo metafísico como contra el subjetivismo, le parecerá una estrategia imposible.

Pero, de hecho, el realismo metafísico y el subjetivismo no son simples «contrarios». En nuestros días tendemos a ser demasiado realistas con respecto a la física y demasiado subjetivistas con respecto a la ética, y estas tendencias están relacionadas. Es *porque* somos demasiado realistas en física, porque consideramos a esta última (o a alguna hipotética física futura) como La Teoría Verdadera, y no simplemente como una descripción racionalmente aceptable, adecuada a ciertos problemas y propósitos, por lo que tendemos a ser subjetivistas con respecto a las descripciones que no podemos «reducir» a la física. Así mismo, llegar a ser menos realista con respecto a la física y menos subjetivista con respecto a la ética son tendencias que también guardan relación.

El argumento del final del capítulo 2 estaba dirigido contra la versión «fisicalista» o naturalista del realismo metafísico. Recordémoslo: supongamos que la interpretación estándar I (bajo la cual «gato» se refiere a gatos, «cereza» a cerezas, etc.) es idéntica o coextensiva con la relación fisicalista R. R se da entre las ocurrencias de «gato» (o los eventos físicos que se dan en alguien que esté usándolas adecuadamente) y los gatos, etc. La interpretación no-estándar J, también descrita, será coextensiva con cierta relación R', definible en términos de R y de los mundos posibles y permutaciones empleadas al construir J (v. Apéndice). De forma que R' se da entre las ocurrencias de «gato» (o los eventos físicos que se dan en alguien que utiliza éstas de la forma estándar) y las *cerezas*, etc. Tanto R como R' son «correspondencias»: bajo ambas son «verdaderas» las mismas oraciones. Las acciones que se exigen para la R'-verdad de una oración (es decir, las acciones que tendrán «éxito» desde el punto de vista del agente) son las mismas que se exigen para que la oración sea R-verdadera. Si R es «idéntica a la referencia»; si R y R', y todas las

demás relaciones que asignan extensiones a nuestras palabras en conformidad con aquellos modos que satisfacen nuestros constreñimientos operacionales y teóricos, no son igualmente correctas, y no lo son porque una de ellas —*R*— es precisamente *la* referencia, entonces este mismo hecho es *inexplicable* bajo una óptica fisicalista.

Este argumento no se dirige sólo contra el realismo metafísico, sino también contra el *reduccionismo*. Si en la imagen fisicalista del mundo no hay nada que dé cuenta del hecho obvio de que «gato» se refiere a gatos y no a cerezas, entonces tenemos una razón decisiva para desestimar la exigencia de que todas las nociones que usamos sean reducibles a términos físicos. Pues no podemos renunciar consistentemente a las nociones de referencia y verdad. Si pienso «Un gato está sobre una estera», entonces me comprometo a creer que «gato» se refiere a algo (si bien no me comprometo con la descripción de la «referencia» del realismo metafísico) y a creer en la verdad de «Un gato está sobre una estera» (si bien tampoco me comprometo con la descripción de la verdad ofrecida por el realismo metafísico).

Una vez repasado el argumento del capítulo 2, veamos ahora cómo afecta a los argumentos en favor del subjetivismo ético. La teoría de la «proyección» nos ofrecía una explicación de la experiencia moral: la experiencia moral es, por así decirlo, un sentimiento subjetivo mal localizado. Contrastemos la teoría de la «proyección» con la siguiente explicación: «Todos los seres humanos tienen un sentido de la justicia y alguna idea de lo bueno. De modo que respondemos (aunque no siempre) a llamamientos tales como "Sea usted amable con los extranjeros, *pues ya sabe usted lo que significa ser extranjero en Egipto*". Nuestra solidaridad se amplía, en parte porque nos hemos persuadido de que *debe* ampliarse; sentimos (a veces) la iniquidad de los actos atroces, aun cuando no encontremos en la víctima a una persona con la que podemos simpatizar fácil o espontáneamente. Encontramos semejanzas entre las injusticias que sufren los demás y las que sufrimos nosotros, y entre los beneficios recibidos por los demás y los que recibimos nosotros. Inventamos palabras morales para las características moralmente relevantes de las situaciones, y comenzamos a establecer gradualmente generalizaciones morales explícitas, que nos conducen a un refinamiento aún mayor de nuestras nociones morales, y así sucesivamente».

A primera vista, esta descripción es más simple y sofisticada que la «teoría de la proyección». (Para empezar, reconoce el papel que desempeñan los *argumentos* en la formación de las actitudes morales.) Sin embargo, muchas personas inteligentes creen que hoy en día debemos rechazar el discurso sobre «un sentido de la justicia» o sobre «la idea de bien» (considerado en un sentido que no es puramente subjetivo), por «anticientífico». De este modo, el conocimiento moral se convierte en algo problemático, quizá completamente imposi-

ble. Pero ¿qué quiere decir aquí «anticientífico»? Creer que hay tal cosa como la justicia no significa creer en *fantasmas,* ni el «sentido de la justicia» es un sentido paranormal que nos permita percibir tales fantasmas. Nadie se propone *añadir* la justicia a la lista de objetos reconocidos por la física, del mismo modo que los químicos del siglo dieciocho proponían añadir el «flogisto» a la lista de objetos reconocidos por la teoría química. La ética *no entra en conflicto* con la física, tal como sugiere el término «anticientífico»; lo único que pasa es que «justo», «bueno» y «sentido de la justicia» son conceptos que pertenecen a un discurso que no es *reducible* al discurso físico. Como hemos visto, hay otros tipos de discurso que no son reducibles a este último, sin que por esta razón sean ilegítimos. El discurso sobre «la justicia», como el discurso sobre la «referencia», puede ser *no*-científico sin ser *anticientífico.*

Para comprender lo que sigue, consideremos cualquier principio básico de la lógica o de la matemática, el principio según el cual la serie de los números enteros puede prolongarse infinitamente («Todo número tiene un sucesor»), por ejemplo, o el principio según el cual un conjunto no vacío de números enteros debe contener un miembro que sea el más pequeño. Supongamos que alguien adelanta la siguiente opinión: «Estos principios son verdaderos para números y conjuntos de números con los que tratamos en la práctica. Por eso llegan a parecer necesarios. Por el mecanismo llamado "proyección", concedemos a los principios mismos este *sentimiento de necesidad,* sentimos que los *enunciados* poseen una misteriosa "necesidad". Pero esta "necesidad" carece de la justificación. Por lo que sabemos, puede que ni siquiera sean verdaderos».

Prácticamente nadie estaría de acuerdo con esta opinión. Casi todos los matemáticos dirían, en lugar de esto, algo como lo siguiente: «La mayoría de los seres humanos tienen, en alguna medida, cierta intuición matemática. De este modo, "vemos" intuitivamente, o bien mediante ejemplos (o mediante un hábil interrogatorio, como el niño-esclavo en el diálogo platónico) que los principios son necesariamente verdaderos».

Kurt Gödel creía que la «intuición matemática» era análoga a la *percepción.* Los objetos matemáticos (a los que denominó «conceptos») están *ahí fuera,* y nuestra intuición nos permite percibir intelectualmente esas entidades platónicas; no obstante, pocos matemáticos se comprometerían con esa metafísica. La comparación gödeliana de la intuición matemática con la percepción revela una idea demasiado simple de esta última. La visión no nos ofrece un acceso directo a un mundo prefabricado, sino que nos ofrece una descripción de los objetos según la cual estos últimos están en parte estructurados y constituidos por la propia visión. Si consideramos que el arco iris del físico es el arco iris «en sí mismo», entonces el arco iris «en sí mismo» no

tiene franjas (el análisis espectroscópico arroja una distribución uniforme de frecuencias); las *franjas* roja, anaranjada, amarilla, verde, azul y violeta son características del *arco iris perceptivo,* no del arco iris físico. El arco iris perceptivo depende de la naturaleza de nuestro propio aparato perceptivo —de nuestra «construcción visual del mundo», según la denominación de Nelson Goodman. (Los «objetos» del físico también dependen de nuestra construcción-del-mundo, como muestra la plétora de versiones radicalmente diferentes que los físicos construyen a partir de los «mismos» objetos.) A pesar de todo, no pensamos que la visión sea *defectuosa* porque se vean franjas en el arco iris; quien *no pudiese* verlas sí tendría una visión defectuosa. La visión está avalada por su capacidad de facilitarnos una descripción que se ajusta a los objetos *para nosotros,* y no a las cosas metafísicas en-sí-mismas. La visión es buena cuando nos permite ver el mundo «tal como es», esto es, un mundo humano y funcional, creado en parte por la propia visión.

Un nuevo axioma propuesto para la teoría de conjuntos, tal como el «axioma de elección», puede adoptarse debido, en parte, a su ajuste a la «intuición» de los matemáticos expertos, y, en parte, a su rendimiento. Si el axioma de elección no produjese resultados que contasen como éxitos matemáticos, el hecho de que algunas personas lo encontrasen «intuitivo» sería de poco interés. La propia intuición matemática se demuestra o se prueba mediante la aprehensión de principios matemáticos y mediante las pruebas subsiguientes. En resumen, la intuición matemática es buena cuando nos permite ver los hechos matemáticos «tal y como son», esto es, como son en un mundo matemático construido por la práctica matemática humana, incluyendo en esta práctica la aplicación de la matemática a otras materias.

Una descripción psicológica o fisiológica de la visión no puede decirnos si la visión de las franjas en el arco iris cuenta como una «visión correcta» o no. Tampoco una descripción psicológica o fisiológica del proceso que acontece en el cerebro cuando «aprehende» el principio de inducción matemática, podría (aún menos) decirnos si el principio es *verdadero* o no. Una vez se admite este punto, no debería sorprendernos que una descripción del proceso cerebral que acontece cuando uno «ve» una acción injusta no pueda decirnos si la acción es realmente injusta.

El discurso sobre la «percepción» moral, como el discurso sobre la intuición matemática, o el discurso sobre la referencia y la comprensión, no es reducible a una imagen fisicalista del mundo. Esto no significa que la física sea «incompleta». La física puede ser «completa», esto es, completa para los propósitos físicos. La física *carece* de la completud de la que carecen todas las teorías, descripciones y discursos particulares. Pues es obvio que ninguna teoría o descripción puede ser completa para *todos* los propósitos. Si la irreducibilidad de

la ética a la física muestra que los valores son proyecciones, entonces los *colores* también lo son. Y también los números naturales. Y, puestas así las cosas, también «el mundo físico». Pero ser una proyección, en este sentido, no es lo mismo que ser algo *subjetivo*[6].

AUTORITARISMO Y PLURALISMO

He estado argumentando que es necesario tener estándares de aceptación racional hasta para tener un mundo, sea un mundo de «hechos empíricos» o un mundo de «hechos de valor» (un mundo en el que hay belleza y tragedia). No hará falta decir que no es posible tener criterios de racionalidad sin aceptarlos, o manteniéndose a distancia de ellos. (La clase de escepticismo consistente en negar que uno tiene algún estándar de racionalidad implica admitir que uno carece de conceptos. Como reconoció Sexto Empírico, ese tipo de empirismo es, en última instancia, inexpresable en el lenguaje.) Tenemos tanto derecho a considerar que algunas inclinaciones «evaluativas» son enfermizas (y todos lo hacemos) como a considerar que lo son algunas inclinaciones «cognitivas». Pero esto no significa rechazar el pluralismo, ni tampoco comprometerse con el autoritarismo.

Ni siquiera en la ciencia, considerar la ciencia como una empresa objetiva (mediante un estándar de «objetividad» que es admitidamente antropocéntrico, pero que es, como observaba Daniel Wiggins, el único estándar del que disponemos), no es mantener que toda pregunta científica tenga una respuesta determinada. Algunas preguntas científicas pueden tener respuestas *objetivamente indeterminadas*, es decir, puede que no se dé una convergencia con respecto a su respuesta, ni siquiera en el límite ideal de la investigación científica; y algunas preguntas científicas pueden tener respuestas determinadas pero relativas al contexto (por ejemplo, «¿Cuál fue la causa del infarto de Juan?», puede tener diferentes respuestas correctas según quien hace la pregunta y por qué la hace). De modo similar, sostener que

[6] En *The Abolition of Man*, McMillan, 1974, C. S. Lewis menciona una versión involuntariamente divertida de la teoría de la proyección. Lewis cita un texto de una escuela secundaria inglesa (cuya identidad se reserva con discreción): «Ustedes recordarán que eran dos los turistas presentes *(Lewis está hablando sobre el conocido relato de Coleridge en la cascada);* que uno la llamó «sublime» y el otro «hermosa»; y que Coleridge se adhirió mentalmente al primer juicio, rechazando el otro con disgusto. Gaius y Titus *(los pseudónimos que Lewis da a los inidentificados autores del texto)* comentan lo que sigue: "Cuando el hombre dice *esto es sublime* parece estar haciendo una observación sobre la cascada... En realidad, su observación no versa sobre la cascada, sino sobre sus propios sentimientos. En realidad, lo que estaba diciendo era *Yo tengo en mi mente sentimientos asociados con la palabra sublime,* o brevemente, yo tengo sentimientos sublimes"».

la investigación ética es objetiva, en el sentido de que algunos «juicios de valor» son definitivamente verdaderos y otros definitivamente falsos, y, más generalmente, que algunas actitudes de valor (y algunas «ideologías») son definitivamente inicuas y que existen algunas que son definitivamente inferiores a otras, no es lo mismo que mantener la posición estúpida de acuerdo con la cual no existen casos indeterminados. (Bernard Williams ha puesto de manifiesto la especial importancia de una clase de casos indeterminados: los casos en los que todas las alternativas son tan atroces que ninguna de ellas sería elegida por una persona idealmente racional y sabia.) Y no hace falta decir que en la ética existen muchos aspectos que son relativos al contexto.

Si hoy discrepamos con Aristóteles es porque somos mucho más pluralistas que él. Aristóteles reconocía que diferentes ideas de *Eudaemonia,* diferentes concepciones del florecimiento humano, podrían ser adecuadas para diferentes individuos, teniendo en cuenta sus diferentes constituciones. Pero parecía pensar que, idealmente, habría algún tipo de constitución que todos deberíamos tener, que en un mundo ideal (olvidando las cuestiones mundanas de quién cultivaría las cosechas y quién cocería el pan) todo el mundo sería filósofo. Nosotros coincidimos con él en que diferentes concepciones del florecimiento humano son adecuadas para individuos con diferentes constituciones, pero vamos más lejos al creer que incluso en el mundo ideal habrían diferentes constituciones, que la diversidad es parte del ideal. Y observamos algún grado de tensión trágica entre los ideales, esto es, que el cumplimiento de algunos ideales excluye siempre el cumplimiento de algunos otros. Pero recalcando el punto de nuevo, creer en un ideal pluralista no es lo mismo que creer que cualquier ideal del florecimiento humano es tan bueno como cualquier otro. Rechazamos algunos de estos ideales como erróneos, como infantiles, como enfermos o como unilaterales.

Tampoco debe confundirse el compromiso con la objetividad ética y lo que es una cuestión muy diferente, el compromiso con el autoritarismo ético o moral. Quizá sea esta confusión la que ha llevado a que un destacado filósofo[7] se adhiera a lo que él mismo considera como una versión limitada del «no-cognitivismo», y a afirmar que «Con respecto a lo que para cada hombre es ''vivir más plenamente'' la autoridad final debe ser el propio hombre». (Notemos la ambigüedad que hay en la expresión «la autoridad final»: ¿quiere decir la autoridad *política* final? ¿Quiere decir autoridad *epistemológica* final? ¿O quiere decir que no existen medios para *ponderar* y *decidir,* tal como sugiere su uso del término «no-cognitivismo»?) El respeto a las

[7] DAVID WIGGINS, «Truth, Convention and the Meaning of Life», *Proceedings of British Academy,* vol. LXII, 1976.

personas como agentes morales autónomos requiere que les reconozcamos el derecho a elegir su punto de vista moral, por muy repulsiva que encontremos su elección. De acuerdo con la filosofía del liberalismo político, también requiere que insistamos en que el gobierno no se asegure de antemano las elecciones morales individuales instaurando una religión o una moralidad de Estado. La oposición incondicional hacia toda forma de autoritarismo político y moral no debería comprometernos con el relativismo o con el escepticismo moral. La razón por la que el gobierno hace *mal* en dictar una moralidad para el ciudadano no es que no existan medios de ponderar qué formas de vida son satisfactorias y qué formas de vida no lo son, o son moralmente inicuas por algún otro motivo. (Si no hubiese algo como la incorrección moral, el gobierno no haría mal en imponer las elecciones morales.) El hecho de que mucha gente tema que si conceden en voz alta algún tipo de objetividad moral encontrarán un gobierno que les haga tragar a la fuerza *su* noción de objetividad moral, constituye sin duda una de las razones por las que tantas personas subscriben un subjetivismo moral que realmente no aprueban.

7. RAZON E HISTORIA

Con el desarrollo científico ha llegado la conciencia de que muchos problemas no pueden ser resueltos mediante los métodos de las ciencias exactas; los ejemplos más obvios los constituyen los problemas éticos e ideológicos. Y con el aumento de nuestra admiración y de nuestro respeto hacia el físico, el cosmólogo, el biólogo molecular, ha llegado una disminución de nuestro respeto y nuestra confianza hacia el pensador político, el moralista, el economista, el músico, el psiquiatra, etc.

En esta situación, algunos han seguido la moda cultural, sosteniendo que en realidad no hay conocimiento fuera de las ciencias exactas (y de las ciencias sociales en la medida en que imitan con éxito a las ciencias exactas, y sólo en esa medida). Esta opinión puede cobrar la forma del positivismo o del materialismo, o de alguna combinación de ambos. Otros han intentado argumentar que la ciencia también es «subjetiva» y arbitraria —ésta es la lectura más popular del exitoso libro de Kuhn, *La estructura de las revoluciones científicas,* pese a que Kuhn afirme ahora que no era esto lo que él proponía. Otros —por ejemplo, los filósofos marxistas y los filósofos de la religión— adoptan una especie de *doble* contabilidad, dejando los problemas técnicos a las ciencias exactas y a la ingeniería, y las cuestiones éticas e ideológicas a un tribunal diferente: el partido, la utopía futura, la iglesia. Pero pocos pueden sentirse cómodos con algunas de estas opciones— con el cientifismo extremo en su forma materialista o positivista, con el subjetivismo y el relativismo, o con alguna de las especies de doble contabilidad. Y si nos enfrentamos con un auténtico problema en este ámbito es precisamente porque nos sentimos incómodos.

El problema es en cierto modo irreal, claro: la misma persona que sostiene que las opiniones éticas y políticas son inverificables, argumenta apasionadamente en favor de *sus* opiniones éticas y políticas. Hume dijo que abandonaba su escepticismo en cuanto abandonaba su estudio; posiblemente los relativistas hagan lo mismo con su relativismo. Pero ello únicamente muestra que nadie puede vivir consistentemente mediante el relativismo; y si esto es todo lo que puede aducirse en contra del relativismo, nos vemos empujados desde éste hacia el existencialismo estilo 1945 («Todo es absurdo, pero hemos de elegir»). ¿Son *tan* distintos?

Para precisar nuestras ideas, recordemos una observación de un

filósofo del siglo pasado, cuyo utilitarismo oculta tras de sí bastante relativismo. Estoy pensando en Bentham y en su desafiante juicio: «prejuicios aparte, el juego de las clavijas tiene el mismo valor que las artes y las ciencias de la música y la poesía». *Prejuicios aparte, el juego de las clavijas es tan bueno como la poesía.*

Esta afirmación es tan traumática para el lector moderno porque entra profundamente en conflicto con nuestros valores actuales. Hemos elevado el arte a un lugar mucho más alto que cualquiera de los que ocupó en los tiempos de Platón o en la Edad Media. Como han sugerido varios autores, para cierta clase de personas cultas el arte es hoy religión, es decir, lo más próximo a la salvación que tenemos disponible.

Bentham está diciendo que la preferencia por «las artes y las ciencias de la música y la poesía» frente al juego infantil de las clavijas es meramente subjetiva, como la preferencia de un helado de vainilla frente a uno de chocolate. Bentham no desea negar que la música y la poesía tengan mayor valor que el juego de las clavijas (una parte importante de la afirmación es «prejuicios aparte»); en el contexto de su utilitarismo, el mismo hecho de que una gran mayoría prefiera la música y la poesía *otorga* a esta última mayor valor. Pero el valor, de ser algo, es *producto* del «prejuicio» (es decir, del interés puramente subjetivo); no hay medios para ponderar el valor relativo del juego de las clavijas y de la poesía, dejando a un lado el hecho de que la gente prefiera esta última. Bentham no está afirmando que preferimos la poesía frente al juego de las clavijas *porque* la poesía tiene más valor que éste, sino más bien lo contrario: que la poesía tiene mayor valor que el juego de las clavijas porque la gente la prefiere. (Aparentemente, *sin ninguna* razón.)

Una exposición tan severa de esta postura la convierte inmediatamente en algo implausible. Por el momento, consideremos una preferencia realmente «subjetiva».

A veces, la gente parece tener en la mente el siguiente modelo de preferencia subjetiva. Existe algún C que es el sabor del helado de chocolate, y existe algún V que es el sabor del helado de vainilla. Existen dos sensaciones G, A, que son «gustar» y «aborrecer». Lo que ocurre, y *todo* lo que ocurre, cuando a Hernández le gusta la vainilla pero aborrece el chocolate y a Fernández le gusta el chocolate pero aborrece la vainilla, es que Hernández experimenta $V + G$ cuando come vainilla y $C + A$ cuando come chocolate, mientras que Fernández experimenta $V + A$ cuando come vainilla y $C + G$ cuando come chocolate.

Sin embargo, tal descripción es psicológicamente ingenua, como Köhler señaló hace ya tiempo. El sabor que tiene la vainilla para Hernández, a quien le gusta la vainilla, no es el mismo sabor que tiene la vainilla para Fernández, quien no la puede soportar. Las cosas son

más bien así: llamemos V_F a la cualidad que tiene la vainilla para Fernández; V_F es un sabor «desagradable»; apenas podemos imaginar que alguien experimente V_F y le guste, e incluso si le gustase, se daría algún tipo de disociación o represión. En resumen, V_F es «intrínsecamente» desagradable, si no metafísica, sí al menos psicológicamente. Y Fernández siente *A* (aborrece el sabor) cuando come vainilla porque la vainilla tiene (para él) la cualidad de sabor V_F. De modo similar, V_H, la cualidad que el sabor de la vainilla tiene para Hernández es intrínsecamente «agradable» (por eso siente *G*). En el lenguaje de G. E. Moore, el sabor V_H y el valor positivo son una unidad orgánica para Hernández, mientras que el sabor V_F y el valor negativo lo son para Fernández. Fenomenológicamente nunca pueden ser separados en dos partes, como sugiere la notación «$V_F + A$» y «$V_H + G$». Casi sin ninguna duda (salvo que entren en juego factores especiales de represión o de disociación) a Fernández también le gustaría el helado de vainilla si provocase en su lengua V_H y no V_F, y Hernández lo aborrecería si provocase en su lengua V_F y no V_H.

Entonces ¿por qué consideramos «subjetiva» la preferencia por la vainilla frente al chocolate? Quiero decir: si no creemos que *todos* los juicios de valor son subjetivos, si no estamos de acuerdo con Bentham en que, prejuicios aparte, el juego de las clavijas es tan bueno como la poesía, ¿por qué sí consideramos subjetiva a esta preferencia? Es obvio que si pensamos que toda preferencia es subjetiva, pensaremos que ésta también lo es; pero lo que nos interesa es el motivo por el cual este juicio no nos parece objetivo, salvo que *seamos* Hernández o Fernández: por qué no tiene el tipo de objetividad que tienen innegablemente muchos otros juicios de valor.

La razón no puede consistir únicamente en la existencia de un desacuerdo. Si creemos que hay juicios de valor objetivos (o justificados), es muy probable que creamos también que algunos juicios de valor discutidos con vehemencia son objetivamente correctos. Los nazis cuestionaron el juicio: *el asesinato indiscriminado de judíos a causa de su filiación racial es inicuo,* pero los antinazis no consideraban que su desacuerdo con los nazis con respecto a este juicio fuese «subjetivo». Aquellos que creen que los homosexuales deberían tener plenos derechos en nuestra sociedad, discrepan de manera violenta de aquellos que piensan que la actividad homosexual o los derechos de los homosexuales deberían proscribirse legalmente. Pero ninguna de las partes en disputa considera que su postura es «subjetiva».

En realidad, es frecuente que el desacuerdo afiance en la gente la convicción de que su posición moral está justificada. De modo que si la discrepancia entre las preferencias de Hernández y Fernández es subjetiva, ello no se debe al hecho de que «algunas personas prefieran el chocolate y otras la vainilla».

Parte del asunto tiene que ver con que la mayoría de la gente no

tiene ninguna preferencia marcada entre la vainilla y el chocolate, pero eso no puede ser lo decisivo. Si la mitad de la población aborreciese el chocolate pero adorase la vainilla, y a la otra mitad le ocurriese lo contrario, seguiríamos considerando (si fuéramos razonables en nuestras preferencias) esta preferencia como una «cuestión de gusto», es decir, subjetiva. Así que la existencia de «neutrales» tampoco es lo decisivo.

En mi opinión, lo decisivo es que sean cuales sean las idiosincrasias biológicas y psicológicas responsables de las preferencias de Hernández y Fernández, no están correlacionadas con características importantes de la mentalidad y del carácter. Si ensayamos el experimento mental de imaginar lo contrario, imaginar que hay un temperamento que consideramos bueno, tanto por sí mismo como por sus efectos sobre los sentimientos, los juicios y las acciones, y otro que consideramos malo, tanto por sí mismo como por sus efectos, y que todo el mundo sabe que el temperamento «bueno» ha manifestado siempre una preferencia por el chocolate, entonces creo que comprobaremos que, en la medida en que logremos representarnos vívidamente este caso como algo real para nosotros, tendremos la impresión de que, en ese mundo, la primera preferencia se consideraría «normal» y «correcta», y la segunda «perversa» y «mostruosa», o algo parecido.

No pretendo afirmar que juzguemos moralmente a *todas* las preferencias según los rasgos del carácter que pensamos que expresan. Algunas «preferencias» son espantosamente importantes en sí mismas: quien pensara que es francamente maravilloso torturar niños por pura diversión, sería recriminado (si hablase en serio) por mantener esa actitud. Pero si se considera que el asunto no es importante en sí mismo. dependerá de lo que creamos que *revela* la preferencia que hagamos de ella un problema o que la tomemos como una «cuestión de gusto». Los juicios de valor a menudo entran en grupos; y estos grupos de juicios de valor revelan con frecuencia ciertos rasgos más o menos permanentes de la mentalidad, de la personalidad y del carácter. Y es la *independencia* de «prefiero el helado de vainilla al de chocolate» con respecto a cualquier «grupo» de este tipo lo que convierte a este juicio en algo «subjetivo» (junto con la ausencia de importancia intrínseca en la misma elección, por supuesto).

Aun cuando la preferencia de Hernández por la vainilla sea «subjetiva», ello no la convierte en irracional o arbitraria. Hernández tiene una razón —la mejor de las razones— para que le guste la vainilla, a saber, *el sabor que tiene para él.* Los valores pueden ser «subjetivos», en el sentido de ser relativos, y con todo ser objetivos; es objetivo que *para* Hernández la vainilla tiene *mejor* sabor que el chocolate. En *The Sovereignty of «Good»,* Irish Murdoch señalaba que filóso-

fos tan diferentes como los existencialistas franceses y los positivistas lógicos compartían de hecho «un modelo común» en relación con el juicio de valor: el modelo según el cual la razón proporciona «hechos» neutrales a la mente, sobre cuya base la voluntad debe elegir «valores» arbitrariamente —y la elección debe ser arbitraria precisamente porque los «hechos» son (por definición) neutrales. Pero ya que la razón no ofrece ningún indicio a la voluntad sobre cómo tomar la decisión (la razón sólo proporciona «hechos», según esta imagen), esta última no tiene ningún *motivo* para su elección arbitraria. Por ello los filósofos franceses la llamaron «absurda», y por eso los filósofos con inclinaciones más naturalistas vieron en el instinto y en la emoción (los sucesores históricos del placer, la categoría-*comodín* de Bentham) las bases últimas de la elección moral.

No obstante, el modelo existencialista-positivista no se ajusta al caso que acabamos de examinar. El «hecho» —*el sabor*— y el «valor» —la exquisitez del sabor— son una misma cosa, por lo menos psicológicamente. Por lo general, las cualidades que experimentamos no son neutrales, y frecuentemente parecen exigir respuestas y actitudes. Uno puede ignorar esas exigencias por una razón válida y suficiente, como cuando un niño aprende a resistir el dolor de una inyección por el beneficio que le proporcionará el agente inmunizante que se le inyecta; pero difícilmente puede negarse que las experiencias particulares sean *prima facie* buenas o malas. (Es bastante significativo que Platón y los medievales ya reconocieran este punto —quizá seamos la primera cultura que concibe la experiencia como algo neutral.)

En el caso *juego de las clavijas-poesía* también se mantiene la no-neutralidad de la experiencia. Nos parece prácticamente imposible imaginar que alguien que realmente aprecie la poesía, alguien que sea capaz de distinguir la auténtica poesía del mero verso, alguien capaz de vibrar con la poesía de talla, *prefiera* un juego infantil frente a artes como la música o la poesía, que enriquecen nuestras vidas. *Tenemos* una razón para preferir la poesía al juego de las clavijas y esa razón descansa en la experiencia que sentimos ante la poesía de talla y en las consecuencias de ésta: el enriquecimiento de nuestra imaginación y de nuestra sensibilidad a través del enriquecimiento de nuestro repertorio de imágenes y metáforas, y la integración de imágenes y metáforas poéticas en nuestras intuiciones y actitudes cotidianas que tienen lugar cuando un poema pervive en nosotros durante varios años. Esas experiencias también son buenas *prima facie* —y no solamente buenas, sino también ennoblecedoras, por utilizar una palabra anticuada.

No obstante, el que puedan haber *razones* para los juicios de valor —razones que sean efectivamente válidas para que determinadas personas emitan determinados juicios de valor— no significa que todos los juicios de valor sean racionales, por supuesto. Los juicios de

valor, juicios hacia los que la gente ha manifestado un apasionado interés, y en cuyo nombre se ha asesinado y torturado, se han basado con demasiada frecuencia en una nociva mezcla de impulsos agresivos e ideas narcisistas. No es sorprendente que cuando un filósofo e historiador relativista como Michel Foucault escribe sobre el pasado, dirija a menudo su atención hacia esas ideas y juicios de valor irracionales. Pero es importante comprender por qué lo hace.

Los escritos de Foucault versan sobre los comienzos de la era moderna (los siglos dieciséis y diecisiete) y, en general, sobre la ideología y la cultura. Su conocimiento de los hechos es legendario, si bien muchos especialistas discrepan de los «hechos» de Foucault. Mientras que algunos de sus libros son sumamente abstractos (por ejemplo *La arqueología del saber*), otros son bastante específicos, por ejemplo *El nacimiento de la clínica* y *Vigilar y castigar*. *El nacimiento de la clínica,* quizá su mejor obra, contiene una parte importante de los argumentos que aparecen en las teorías más abstractas de Foucault.

Lo que Foucault intenta mostrar es que la «clínica», es decir, la institución del hospital y las instituciones médicas adyacentes, fue tanto el resultado del aumento del conocimiento y de la técnica científica como el reflejo de cierta ideología con respecto a la salud y a la enfermedad. Esta ideología estuvo a su vez asociada con cambios ideológicos más amplios, especialmente con el nacimiento del individualismo del siglo diecisiete. Y sugiere que la «clínica» no es la forma óptima de tratar a la mayoría de los pacientes, y que nuestra creencia en que sí lo es no es más que un tipo de prejuicio ideológico, una especie de locura.

Aquí se vislumbra una insinuación de mayor alcance: las convicciones ideológicas y los juicios de valor asociados a éstas son cuestiones bastante arbitrarias [1]. En materia de ideología no hay una posición objetiva en la que situarse (si exceptuamos, claro está, el misterioso punto de vista de *La arqueología del saber* de Foucault, pretendidamente objetiva).

Para comprender lo que Foucault quiere decir, consideremos un ejemplo más familiar y menos controvertido. En la Edad Media se creía que la monarquía era la forma de gobierno natural y adecuada. Esta creencia se basaba en parte en otras creencias fácticas que hoy

[1] Y también se insinúa que nuestro pensamiento está determinado por el lenguaje que usamos. Foucault habla de «sistema implícitos que determinan, sin que lo sepamos, la mayor parte de nuestros comportamientos más familiares» (véase J. SIMON, «A Conversation with Michel Foucault», *Partisan Review,* n.° 2, 1971, p. 201). El estructuralismo francés, al menos el que representaban Foucault, Althusser, Lacan, Deleuze, etc., parece equivaler frecuentemente a (1) determinismo; (2) relativismo; (3) la pretensión de que el estructuralismo es una «ciencia lingüística».

consideramos injustificadas (por ejemplo, la creencia en que la democracia conduce inevitablemente a la anarquía y a la tiranía) y en parte en la autoridad de la Iglesia. La opinión de la Iglesia se basaba en parte en consideraciones políticas (la Iglesia era la religión del Estado), pero este hecho pasaba inadvertido, debido a que se creía que Dios la inspiraba y la nombraba intérprete de su palabra y de su voluntad. Foucault está insinuando que las creencias que se mantenían en el pasado reciente y, por implicación, las que hoy mismo mantenemos, no son más racionales que la creencia medieval en el derecho divino de los reyes.

Consideremos por un momento esta última creencia. Si no pensamos que haya una *buena razón* para creer en la existencia de un Dios personal, que dispone que tengamos ciertas formas de vida y ciertas estructuras sociales, esta creencia será tildada de irracional (lo cual no significa negar que la creencia responda a auténticas necesidades psicológicas). Y aun cuando creamos en Dios, si no creemos que la iglesia tiene acceso especial a sus deseos, pensaremos que la doctrina del derecho divino de los reyes era y es una doctrina irracional. Y finalmente, aun cuando seamos católicos, no aceptaremos de buen grado el que el apoyo de la iglesia a la monarquía se base más en consideraciones políticas que en la revelación o en una sólida teología. En resumen, la creencia en el derecho divino de los reyes carece y siempre careció de una base racional adecuada.

Entonces, ¿cómo surgió esa creencia? La respuesta habitual apelaría, en parte, a factores políticos y económicos (no hace falta ser marxista para aceptar que estos factores están entre los determinantes de la teología), y, en parte, a factores psicológicos. El consuelo que proporciona la creencia en un Dios personal y en otra vida es evidente, como lo es el consuelo que obtiene el creyente en una iglesia infalible y en un orden social divinamente fijado. En resumen, la autogratificación narcisista y el condicionamiento social fueron los auténticos determinantes de esta creencia. Y si esta creencia es realmente *típica,* si en realidad es representativa de *todas* nuestras creencias ideológicas, entonces tales factores son los auténticos determinantes de *toda* ideología.

Y si tantos pensadores franceses tienen en gran estima a Marx, Freud y Nietzsche, ello es debido a que profesan una creencia de esta índole. Marx, Freud y Nietzsche tienen en común lo que sigue: consideran a las religiones y a las ideas éticas que abrigamos como reflejos de lo *irracional,* del interés de clase (en el caso de Marx), del inconsciente (Freud y Nietzsche) o de la voluntad de poder (Nietzsche)[2].

---

[2] No estoy acusando a Marx, a Freud o a Nietzsche de extraer consecuencias relativistas a partir de esto.

Por debajo de lo que nos complacemos en considerar como nuestras intuiciones espirituales y morales más profundas, se encuentra un hervidero de impulsos de poder, intereses económicos y fantasías egoístas. Esta opinión es hoy el mordiente del relativismo.

Al mismo tiempo, ningún relativista podrá utilizar el término «irracional» como yo acabo de hacerlo al describir el punto de vista relativista. El propio relativismo excluye esta utilización.

Cuando enseñé estas páginas a un amigo relativista, se indignó por mis afirmaciones sobre el derecho divino de los reyes. ¿Acaso no estaba yo enterado de que muchos hombres inteligentes habían sido convencidos mediante argumentos filosóficos de que esta doctrina era correcta? ¿Acaso estaba ofreciendo una explicación marxista o freudiana? La creencia en el derecho divino de los reyes era *racional,* ¡faltaría más!

Mi réplica consistía en que bien puede haber un sentido de «racional» en el que lo sea cualquier opinión que pueda defenderse con inteligencia y persuasión partiendo de los presupuestos compartidos por una cultura, pero *ese* sentido no puede ser el único, ni siquiera el sentido que tenga mayor importancia normativa. Los judíos aceptaron a Moisés como legislador y profeta porque su doctrina cubría auténticas *necesidades* religiosas, culturales y nacionales; y esto no es lo mismo que ser convencido mediante un argumento racional. Más tarde, los profetas ungieron a los reyes judíos (después de intentar disuadir a los judíos de que tuvieran reyes), lo que difícilmente prueba que los últimos reyes cristianos fueran nombrados por la divinidad. El cristianismo, que compartía la biblia judía, se convirtió en la religión del Imperio Romano, y es difícil creer que esto se debiese a que los ciudadanos o el emperador tuvieran una prueba racional de que el cristianismo era verdadero. Los emperadores romanos fueron ungidos a partir de ese momento (como lo habían sido los reyes judíos), lo que difícilmente prueba que fuesen nombrados por la divinidad. Finalmente, una vez aceptadas las premisas del cristianismo, y *partiendo de éstas,* podían ofrecerse «argumentos racionales» en favor del derecho divino de los reyes. Pero exponer esto diciendo que en el Bajo Imperio Romano, o en la Edad Media, «La creencia en el derecho divino de los reyes era perfectamente racional», es degradar la noción de racionalidad.

Hegel, quien introdujo la idea de que la Razón varía en la historia, empleó dos nociones de racionalidad: hay un sentido en el cual lo racional se mide por el nivel alcanzado por el Espíritu en un momento dado del proceso histórico; hay quien pretende que la creencia en el derecho divino de los reyes fue racional en su tiempo en un sentido análogo. Pero también hay una noción límite de racionalidad en el sistema de Hegel: la idea de aquello cuyo destino es ser estable, la autoconciencia del espíritu que ya no será transcendida. Cuando los

relativistas de nuestros días «naturalizan» a Hegel, sacrificando el concepto-límite de verdadera racionalidad, transforman su doctrina en un relativismo cultural que se trunca a sí mismo.

No obstante, ningún relativista quiere ser relativista con respecto a *todo*. ¿Cómo imponen límites a su propio relativismo estos pensadores franceses? La respuesta varía con el pensador. En el caso de un marxista como Althusser, la respuesta adoptada es una versión de la teoría del «interés de clase»: todas las «ideologías» son producto de factores no racionales, pero aquellas ideologías que son producto de los intereses de la clase trabajadora (en la era presente) son «justas» y propenden hacia la liberación humana, mientras que aquellas ideologías que brotan de los intereses de la clase explotadora son «injustas» y producen miseria. Sin embargo, Althusser se desmarca de aquellos que expusieron con anterioridad este punto de vista relativista-de-clase, resistiéndose a afirmar que la ideología marxista (la ideología de la «clase trabajadora») es *verdadera,* o *está más próxima a la verdad* que la ideología burguesa. Las ideologías pueden ser «justas» o «injustas», pero no verdaderas o falsas[3]. («Verdadero» y «falso» sólo se aplican, según Althusser, en el «laboratorio científico» y, presumiblemente, sólo a aquellos enunciados que tienen condiciones empíricas de prueba.) En sus obras más recientes, Foucault también parece decantarse hacia un relativismo-de-clase, aunque es difícil estar seguro de ello. Lo significativo de este punto de vista, al menos en su forma althusseriana y radical, es su esfuerzo por preservar la afirmación relativista radical de que ninguna «ideología» puede ser racional, si bien conservando la idea de que algunas ideologías (la ideología predilecta —en el caso de Althusser, el marxismo-leninismo) pueden ser buenas, y distinguiendo entre ideologías «buenas» y «malas» y «justas» o «injustas» sobre una base distinta de la aceptabilidad racional. La idea es que aunque toda ideología es adoptada por causas irracionales o no-racionales, algunas de éstas (las que defienden los intereses de la clase trabajadora) son buenas, y producen buenas ideologías (por definición), mientras que otras son malas y producen malas ideologías. En lugar de juzgar las ideologías por sus *razones* (que son siempre *racionalizaciones*) hemos de juzgarlas por sus causas.

No obstante, este medio de limitar el relativismo de uno mismo

[3] De acuerdo con Althusser «las proposiciones filosóficas son tesis». «Las tesis filosóficas pueden mantenerse *negativamente* como proposiciones dogmáticas, en cuanto que no son susceptibles de demostración en el *estricto* sentido científico del término (el sentido que tiene «demostración» en la lógica o la matemática) ni de prueba (en el sentido de «prueba» en las ciencias experimentales)... Ya que no pueden ser demostradas ni probadas científicamente, no puede decirse que las tesis filosóficas sean «verdaderas» (demostradas o probadas, como en la física y en la matemática), sólo puede decirse que son "justas"», *Philosophie et Philosophie Spontanee des Savants,* pp. 13-14, Maspero, 1967.

es impracticable. Pues, ¿en qué se basa el juicio de que la victoria de los «intereses de la clase obrera» traerá consigo un mundo libre de guerras y de racismos, y no un totalitarismo imperialista disfrazado de «socialismo»? Si la respuesta consiste en afirmar que este último no sería un «verdadero» socialismo, o que no reflejaría los verdaderos intereses de la clase obrera, entonces ¿en qué se basa el juicio que afirma que alguna institución (por ejemplo, el partido comunista francés, del cual Althusser es un dirigente) o algún determinado programa político, promoverá los «verdaderos» intereses de la clase obrera y el «verdadero» socialismo? Si estas creencias son racionalmente justificables, entonces no toda ideología es irracional; si no lo son, entonces la pretensión de que alguna institución determinada o algún determinado programa político es «justo» debe ser tan absolutamente irracional como se afirma que lo es cualquier otra pretensión «ideológica». Si todo el pensamiento humano acerca de cuestiones ideológicas es una locura autocomplaciente, entonces el pensamiento en torno a cuáles son las creencias que brotan de los «intereses de la clase obrera» y cuáles no lo son, también debe serlo.

No obstante, cuando volvemos a Foucault observamos, si ignoramos los síntomas de su obra más reciente que muestran que también él se está radicalizando, que su motivo para dirigir la atención a donde lo hace es *precisamente* sugerir el carácter absolutamente no racional (y de hecho, irracional) de las auténticas *razones* que tiene la gente para adoptar posturas ideológicas. Y la noción de ideología abarca aquí muchas cosas; no sólo están siendo discutidos el comunismo, el fascismo, la democracia, el derecho divino de los reyes, etc. La creencia de que alguien está «enfermo» y necesita una «cura», la creencia de que alguien es un «criminal» y ha de ser «rehabilitado», si es posible, y muchas de nuestras creencias cotidianas, son «ideologías», según el sentido que le dan tales pensadores a ese término. En realidad, para el ojo avizor del sociólogo marxista o del filósofo francés, casi *todas* las creencias son «ideológicas». Quizá «Si dejo caer este vaso, entonces se romperá» sea una creencia ideológicamente neutral, pero apenas queda otra.

Puede dar la impresión de que he pasado por alto el punto esencial de la argumentación de Foucault. El diría que lo importante no es, en modo alguno, que las perspectivas ideológicas del pasado fueran absurdas e irracionales, sino más bien que toda ideología, en el amplio sentido que le da a ese término, e incluyendo a nuestra ideología actual, es relativa a la cultura. Foucault intenta mostrarnos que toda cultura vive, piensa, observa y hace el amor mediante un conjunto de presuposiciones directivas cuyos determinantes son no-racionales. Si las ideologías precedentes nos parecen hoy «irracionales» es porque las juzgamos a partir de nuestra noción de racionalidad, limitada culturalmente.

Lo inquietante de la explicación de Foucault es que los determinantes que tanto él como otros pensadores franceses desenmascararon, son *irracionales según el estado actual de nuestros conocimientos.* Si nuestra ideología actual es el producto de fuerzas que son incoherentes *según sus propios conocimientos,* entonces es internamente incoherente. Los pensadores franceses no son *meramente* relativistas culturales; están atacando nuestra noción de racionalidad desde dentro, y esto es lo que inquieta al lector.

El relativismo cultural no es, en absoluto, una doctrina nueva. Desde que ha habido antropólogos, éstos nos han predicado el relativismo cultural. Pero sería un error asimilar el relativismo de Foucault con el relativismo de antaño.

Cuando un antropólogo nos predica el relativismo, normalmente cita prácticas y creencias exóticas que en principio se nos antojan irracionales o repulsivas o ambas cosas, y continúa mostrando que éstas promueven en realidad el bienestar y la cohesión social. En resumen, muestra (en la medida que el ejemplo sea razonable) que lo que en nuestra sociedad se considera como algo incorrecto o irracional puede ser por el contrario correcto y «razonable» en otras circunstancias naturales y sociales.

Ahora bien, los antropólogos extraen a menudo conclusiones erróneas a partir de sus propios ejemplos (y algunos son bastante menos claros de lo que ellos piensan). Un antropólogo afirmará con demasiada frecuencia que «Todo es relativo», queriendo decir que no hay ningún medio de ponderar y decidir lo que es correcto y lo que es erróneo. Richard Boyd me ha sugerido que, muy a menudo, el motivo es político: convencernos de que detengamos la destrucción de las culturas primitivas, atacando nuestra creencia en la superioridad racional y moral de la cultura a la que pertenecemos. Por desgracia, su argumento es demasiado confuso. Los ejemplos que ofrece el antropólogo (cuando son buenos) muestran que lo que sea correcto o erróneo es relativo a las circunstancias, y *no* que no exista ni la corrección ni la incorrección, ni siquiera en las circunstancias que se especifiquen. Su argumento contra el relativismo cultural equivale a lo que sigue: las demás culturas no son objetivamente peores que la nuestra (ya que según él, no hay algo que sea objetivamente mejor o peor), por consiguiente, *son exactamente tan buenas* como la nuestra; por lo tanto es un error destruirlas.

Este argumento yerra. La conclusión exige que «son exactamente tan buenas como» signifique objetivamente *exactamente tan buenas como* (al menos conforme a nuestros conocimientos); pero de la inexistencia de valores objetivos no puede seguirse que cualquier cosa sea exactamente tan buena como cualquier otra (en el sentido exigido), sino que no existe la relación de «ser exactamente tan bueno como». Si los valores *fueran* en efecto arbitrarios, ¿por qué motivo

no deberíamos destruir cualquier cultura que nos plazca destruir?

Por fortuna, existen mejores argumentos para criticar el imperialismo cultural que aquél que niega todo valor objetivo. El antropólogo puede tener un buen motivo, pero se ha equivocado de argumento. También se equivoca en el término «ser relativo». En realidad, lo que sus ejemplos confirman es el «relativismo objetivo» de Dewey. Ciertas cosas son correctas —*objetivamente correctas*— en ciertas circunstancias, e incorrectas —*objetivamente incorrectas*— en otras, y la cultura y el entorno constituyen las circunstancias relevantes. Aunque el antropólogo esté en lo cierto con respecto a esto último, ello no significa que los valores sean «relativos», en el sentido de ser meras cuestiones de opinión o de gusto.

Una vez liberado de sus confusiones conceptuales, el argumento del antropólogo no debe inquietarnos. Hemos de dar la bienvenida a sus observaciones, pues tienden a aumentar nuestra sensibilidad y a atacar nuestra presunción de superioridad cultural. Pero la simple comparación del argumento de Foucault con el del antropólogo nos brinda ya la diferencia: Foucault no sostiene que las prácticas del pasado fueran *más* racionales de lo que parecen, sino que todas las prácticas son *menos* racionales: en realidad, sus principales determinantes son la sinrazón y el poder egoísta. Sólo hay una semejanza superficial entre esta doctrina y el antiguo relativismo cultural.

El hecho es que la postura que hemos estado describiendo se encuentra al servicio de una tentación intelectual resultante del aumento de nuestro conocimiento y de nuestra sensibilidad con respecto a los mecanismos sociológicos y psicológicos. Tanto ese conocimiento como esa sensibilidad son en parte pretendidos y en parte reales; la tentación es caer en la trampa de concluir que todo argumento racional es una mera racionalización, y, a pesar de eso, seguir intentando *argumentar racionalmente* en favor de esta posición.

Si *todo* «argumento racional» fuera una mera racionalización, entonces no sólo no tendría sentido intentar argumentar racionalmente en favor de cualquier punto de vista, sino que ni siquiera tendría sentido *mantener* un punto de vista. Si considero que mi propio asentimiento y mi propio disentimiento son conductas caprichosas, entonces debo dejar de asentir y de disentir —algo para lo que no puede haber asentimiento o disentimiento racional, sino sólo una disparatada parodia de discusión racional, no puede llamarse un *enunciado*. Como Sexto Empírico, quien acabó por concluir que su propio relativismo no podía expresarse en un enunciado (ya que ni siquiera podría saber el enunciado «No sé»), el relativista moderno, si fuera consistente (¿y *cómo* podría mantener *consistentemente* una doctrina que convierte la noción de consistencia en un sinsentido?) debería acabar considerando sus propias proferencias como meras expresiones emotivas.

Afirmar esto no es negar que podamos pensar racional y correctamente que *algunas* de nuestras creencias son irracionales, sino afirmar que la insistencia en nuestra maldición intelectual no puede traspasar ciertos límites, so pena de convertirse en algo ininteligible. Por ejemplo, nosotros *discutimos* doctrinas como las propuestas por Foucault; nos esforzamos en ser imparciales, intentamos adoptar lo que Popper llama una «actitud crítica», y buscamos con ahinco evidencias y argumentos que podríamos haber pasado por alto, aun cuando atenten contra nuestros propios puntos de vista. Nada de esto tendría el más mínimo sentido si no pensáramos que la práctica de la discusión y de la comunicación, y las virtudes de la crítica y de la imparcialidad, tienden a extirpar las creencias irracionales —si no de inmediato, sí al menos gradualmente, con el tiempo— y a mejorar la afirmabilidad fundada de nuestras conclusiones. Pese a que la racionalidad no puede ser definida mediante un *«canon»* o conjunto de principios, *sí tenemos una concepción en evolución de las virtudes cognitivas, que nos sirve de guía.*

Habrá quien objete que esta concepción «no nos lleva muy lejos». Rudolf Carnap y John Cardinal Newman fueron pensadores responsables y cautos, y ambos se comprometieron con las virtudes cognitivas que acabamos de mencionar, pero nadie cree que si hubieran vivido en la misma época y hubiesen estado dispuestos al diálogo uno de ellos hubiese convencido al otro. Pero el hecho de que no haya un medio universalmente satisfactorio para resolver toda disputa no muestra que en cada caso no haya concepciones mejores y peores. La mayoría de nosotros pensamos que el catolicismo de Newman era un tanto obsesivo; y la mayoría de los filósofos creen que Carnap, pese a ser un filósofo brillante, empleó muchos argumentos endebles. El que emitamos estos juicios muestra que tenemos una idea regulativa de lo que debe ser un intelecto justo, despierto y equilibrado, y que creemos que hay medios para ponderar cómo y por qué ciertos pensadores no alcanzan ese ideal. Algunos replicarán: «¡Y qué!; a la hora de resolver una disputa real, no nos encontramos en una posición más cómoda ahora que cuando no existía *ninguna* noción de aceptabilidad externa a los puntos de vista en conflicto a la cual apelar». Y ello es cierto cuando se trata de una disputa irresoluble, como la disputa que acabamos de imaginar entre Carnap y Newman; pero no es cierto que a la larga estemos en la misma situación si abandonamos la idea de que en realidad hay algo como la imparcialidad, la consistencia y la razonabilidad, aun cuando en la práctica sólo logremos aproximarnos a ellas— o si admitimos la opinión de que, con respecto a estas cuestiones, sólo hay creencias subjetivas y ningún medio que cribe cuáles de estas «creencias subjetivas» son correctas.

La analogía que (ocasionalmente) he esbozado entre la discusión política y la discusión filosófica quizá pueda servirnos de ayuda. Uno

de mis colegas es un conocido defensor de la opinión según la cual todo gasto público en «bienestar social» es moralmente impermisible. En su opinión, hasta el sistema de escuela pública es una aberración moral. Si fuese abolido el sistema de escuela pública, junto con la ley de educación obligatoria —ley que creo que también considera como una impermisible interferencia del gobierno en la libertad individual—, entonces las familias más humildes no podrían enviar a sus hijos a la escuela, y optarían por dejar que sus hijos se convirtieran en adultos analfabetos; pero según él, este problema debe resolverlo la caridad privada. Sería dañino que la gente no fuese lo suficientemente caritativa como para prevenir el analfabetismo masivo (o la inanición generalizada de los ancianos), mas esto no legitima la acción del gobierno.

En *mi* opinión, *sus* premisas fundamentales —que el derecho de propiedad sea algo absoluto, por ejemplo— son contraintuitivas y no están respaldadas por un argumento suficiente. En *su* opinión, yo me encuentro bajo la tiranía de una filosofía «paternalista» que él considera insensible como respecto a los derechos individuales. He aquí un desacuerdo extremo, y más bien un desacuerdo en «filosofía política» que un «desacuerdo meramente político». Y aunque muchos desacuerdos políticos entrañan desacuerdos en filosofía política, rara vez son tan radicales como éste.

¿Qué ocurre con tales desacuerdos? Cuando ambas partes defienden su causa con inteligencia, a veces todo lo más que puede pasar es que se diagnostique y se delimite el origen del desacuerdo. A menudo, cuando el desacuerdo es menos fundamental que el que he descrito, ambas partes pueden modificar su punto de vista en mayor o menor medida. Si no se logra un acuerdo real, quizá puedan establecerse compromisos más o menos aceptables para una u otra de las partes.

Por desgracia, la discusión política inteligente entre gente que mantiene distintas perspectivas es rara hoy en día; pero cuando se da, es de lo más gratificante. La actitud que mantenemos ante nuestro rival en la discusión constituye una mezcla interesante. Por una parte, reconocemos y apreciamos ciertas virtudes intelectuales de la mayor importancia: apertura mental, voluntad de considerar razones y argumentos, capacidad de aceptar las críticas certeras, etc. Pero ¿qué ocurre con los aspectos fundamentales en los que no podemos *estar de acuerdo*? Sería bastante deshonesto fingir creer que *en este punto* no existen razones mejores o peores. Yo no creo que sea una mera cuestión de gusto considerar que la obligación de la comunidad de tratar caritativamente a sus miembros prevalece sobre los derechos de propiedad; tampoco lo cree mi oponente. Cada uno de los dos consideramos que el otro carece de cierto tipo de sensibilidad en este punto. Para ser absolutamente honestos, ambos sentimos algo semejante a un *desprecio,* no hacia la inteligencia del otro —pues tenemos en la

más alta estima la inteligencia del oponente— ni hacia su *persona* —ya que tengo el mismo respeto hacia la honestidad, integridad, amabilidad de mi colega que hacia muchas personas que comparten mis opiniones políticas «liberales»—, sino hacia cierto complejo de emociones y juicios que encontramos en el rival.

Pero ¿no estoy siendo en este punto algo menos que honesto? Digo que respeto la inteligencia de Bob Nozick y, sin duda, la respeto. Digo que respeto su carácter, y sin duda lo hago. Pero si siento desprecio (o algo de esta índole) hacia cierto complejo de emociones y juicios que encuentro en él, ¿no es eso sentir desprecio hacia *él?*

Es doloroso examinar, y normalmente la cortesía nos impide hacerlo con justicia, cuáles son las actitudes que mantenemos hacia aquellos que apreciamos, pero con los que discrepamos en algunos puntos. El hecho es que nadie que sea del todo maduro aprecia y respeta *todo* con respecto a *cualquier persona* (y menos aún con respecto a uno mismo). No hay contradicción entre tener aprecio y respeto por alguien y, a pesar de todo, hallar en él algo parecido a una debilidad intelectual y moral, del mismo modo que no la hay entre respetarse y apreciarse a sí mismo y, aún así, reconocer en uno mismo algo parecido a una debilidad intelectual y moral (o emocional, etc.)

Quiero insistir en que existe toda la diferencia del mundo entre un oponente que posea las virtudes intelectuales fundamentales de apertura mental, respeto por la razón y autocrítica, y uno que no las posea; entre un oponente con un impresionante y pertinente bagaje de conocimiento objetivo y uno que no lo tiene; entre un oponente que meramente desahogue sus sentimientos y sus fantasías (que es lo que suele hacer la gente en lo que pasan por ser discusiones políticas) y un oponente que razone meticulosamente. Y la actitud ambivalente de respeto-desprecio es honesta: respeto por las virtudes intelectuales del otro, desprecio por sus debilidades intelectuales o morales (de acuerdo con nuestros propios conocimientos, desde luego, pues siempre partimos de éstos). El «respeto-desprecio» puede parecernos algo casi *deshonesto* (especialmente si se confunde con el respeto despectivo, que es algo completamente distinto). Y lo sería si nuestro desprecio fuese por el otro como persona, y no sólo por el complejo de juicios y sentimientos que encontramos en él. Pero es una actitud mucho más honesta que el *falso relativismo;* esto es, que la pretensión de que no se pueden ofrecer razones, ni siquiera razones mejores o peores, con respecto a un asunto, cuando en realidad, para nuestros fueros, encontramos una opinión razonable y la otra irracional.

Puede ser útil descender del nivel abstracto en el que hemos estado discutiendo y considerar un ejemplo relativamente simple. Consideremos un juicio que la mayoría de la gente de la calle muchas veces está dispuesta a admitir: que la paz es preferible a la guerra. (Foucault nunca discute estos juicios, por la misma razón que Swift nunca

los describe: ambos son escritores *satíricos*. Sólo les interesa la locura de la sociedad y no su sensatez —cuando ésta existe.) No hay duda de cuál es el origen de tal juicio. Estamos demasiado familiarizados con los horrores de la guerra, con lo que ésta provoca a niños y adultos, a combatientes y a no combatientes, al propio suelo y a la vegetación. Y aun cuando este juicio brote en parte del autointerés, ello *no* lo hace irracional, sino todo lo contrario.

Aún así, poblaciones enteras pueden adherirse al juicio opuesto: que la guerra es preferible a la paz, y no por razones de legítima defensa. La agresividad y la fascinación pueden avivar en la gente una sed nacional de sangre. Pero, una vez más, lo que esto muestra no es que *todos* los juicios de valor sean irracionales, sino sólo que algunos sí lo son: y que es muy difícil distinguirlos cuando uno no está dispuesto a dejar de lado su partidismo y a criticar sus propias creencias (y por ello concedemos tanta importancia a la imparcialidad y a la actitud crítica entre las virtudes cognitivas).

Que algunos juicios de valor sean racionales y objetivos no quiere decir que nuestro discurso abstracto acerca del capitalismo, de la democracia, del socialismo, de los derechos, de la autonomía, etc., no sea con frecuencia un sinsentido. Aun cuando lo que queremos decir acerca de alguna cuestión general sea correcto, con frecuencia tenemos problemas para expresarlo adecuadamente, sobre todo si no estamos habituados a la expresión de ideas abstractas. El caso del antropólogo que afirma que no hay valores objetivos, cuando lo que pretende decir es que los valores son relativos a las circunstancias, es un caso característico. Aunque tengamos éxito al expresar lo que efectivamente queremos decir, *hay* intensas fuerzas de tipo no-racional que tienden a influir en nuestros juicios. En este punto, mi propósito no es negar que el poder puede corromper y el narcisismo puede seducir nuestros juicios; es negar que estemos desvalidos ante estas intensas fuerzas, tan desvalidos que sería ocioso (y de hecho, decepcionante) intentar juzgar con inteligencia y justicia. Afirmar que *podemos* ser racionales no es afirmar que podemos ser infalibles. Por el contrario, tal como señala Irish Murdoch, el esfuerzo por conseguir una postura razonada y racional es en esencia algo progresivo e «infinitamente perfectible» [4].

Un relativista inteligente no tendría problema alguno en aceptar lo que he venido diciendo. Un relativista no necesita ocuparse en socavar la racionalidad de todos los juicios de «valor», ni en defender la imagen que Foucault defiende en relación con la historia —la historia como una serie discontinua de «discursos» o «ideologías» que tienen éxito o no por alguna razón no-racional. Un relativista más mo-

[4] *The Sovereignty of Good*, p. 23.

desto podría contentarse con estar de acuerdo con Dewey[5] en que algunos valores son objetivamente relativos —es decir, que son racionales *dadas las circunstancias,* la naturaleza y la historia de los que los crean. Un relativista modesto mantendría que lo que importa es precisamente la *relatividad* de todos los valores. La «objetividad» que niega no es la «objetividad» que Dewey afirma —la objetividad de cualquier juicio que esté justificado en su marco existencial efectivo—, sino más bien la objetividad que Platón afirmaría, la objetividad espuria (diría el relativista) que pretende hablar desde un punto de vista absoluto, aparte *de* toda circunstancia y válido *para* toda circunstancia.

Si no nos satisface aceptar este relativismo modesto, si nos inquietan los propios escritos éticos de Dewey, no creo que se deba a que suspiremos por los Absolutos. Cuando pretendo que el asesinato y el sufrimiento de la gente inocente es inicuo, en realidad no me preocupa la cuestión de si este juicio sería válido para un ser con una constitución y una psicología *totalmente* distinta a la nuestra. Si hay seres en Alpha Centauro, por ejemplo, que no pueden sentir dolor, y que tampoco conceden demasiada importancia a la muerte de los individuos, entonces, con toda probabilidad, nuestra protesta ante «el asesinato y el sufrimiento» les parecerá mucho ruido y pocas nueces. Pero el mismo hecho de que tal forma de vida sea absolutamente distinta a la nuestra quiere decir que estos seres no pueden comprender las cuestiones morales que tal protesta trae consigo. Y aunque nuestra «objetividad» sea objetividad hablando desde un punto de vista humano, aun así es una objetividad suficiente.

Lo preocupante es que la doctrina del «relativismo objetivo» de Dewey no puede habérselas con el caso de los nazis (si bien Dewey sí lo habría discutido). Deseamos afirmar que las metas de los nazis son sumamente atroces, y la afirmación: «Esto es cierto en relación a tus intereses, pero falso en relación a los intereses de los nazis», constituye precisamente el tipo de «relativismo moral» que encontramos repulsivo. El relativismo objetivo parece ser la doctrina correcta para muchos casos morales, pero no para aquellos en los que los derechos y las obligaciones son nítidos y manifiestos, y nos parece que hemos de elegir entre lo correcto y lo incorrecto, entre el bien y el mal.

En realidad, la noción moderna e instrumental de la racionalidad es «relativista objetiva» hasta cierto punto. El núcleo de esta noción es una dicotomía engañosamente simple: la idea es que la elección de fines o metas no es *ni racional ni irracional* (los requisitos se satisfa-

[5] Véase «Dewey's Theory of Valuation», en *The Encyclopedia of Unified Science,* vol. II, n.° 4, University of Chicago, 1939.

cen, siempre que se dé una mínima coherencia), mientras que la elección de medios *es racional en la medida en que es eficiente.* La racionalidad es un predicado que atañe a los medios, y está fundido por entero con la eficiencia. De modo que la preferencia de Hernández por el helado de vainilla frente al de chocolate no es ni racional ni irracional, pero la de elegir vainilla y no chocolate en una ocasión determinada sería racional para Hernández, dado su «orden de preferencias». Por lo general, los modernos sociólogos suponen que esta concepción (que se retrotrae al *dictum* humeano: «La razón es y debe ser la esclava de las pasiones», y que, en el fondo, está bajo la influencia de Bentham) es la concepción correcta. Esta concepción ha desempeñado su papel en la economía del bienestar y en muchas otras áreas. En economía, la noción moderna de *Pareto Optimum* es un intento de conseguir una noción de condiciones económicas óptimas que tenga en cuenta tan sólo la eficiencia de los medios, y no implique ningún «juicio de valor» con respecto a las metas de los diversos agentes económicos; y si esta noción es de interés contemporáneo, es precisamente *porque* se supone que la elección de los medios está sujeta a la crítica racional, mientras que la elección de los fines no lo está.

No obstante, toda esta concepción pierde gran parte de su persuasiva atracción cuando comprendemos que descansa sobre una teoría psicológica excesivamente simplificada. En el esquema benthamita, los fines, las metas y las preferencias se tratan o como parámetros individuales fijos (esto es, el aprendizaje individual se representa como un proceso consistente en aprender a estimar mejor las consecuencias y probables consecuencias de las acciones, y a alcanzar los fines de una manera más eficiente, pero no como un proceso de adquisición de nuevos fines) o como parámetros individuales que, si no son *fijos,* sólo cambian como resultado de factores que no tienen un *status* racional, de los que el teórico no puede dar cuenta.

Bernard Williams[6] ha señalado que los fines individuales, y no sólo los medios para alcanzarlos, pueden ser criticados racionalmente de varias maneras, que se harán patentes una vez traspasemos la estrecha psicología benthamita.

La concepción «benthamita» no admite que se dé *un* caso en el que un individuo pueda ser persuadido para que abandone una meta (o, en cualquier caso, para que abandone la *persecución* de una meta) mediante una crítica racional: lo que ocurre en este caso, es que el individuo ha estimado defectuosamente las consecuencias de la acción, y el resultado es una grave subestimación de los *costos* implícitos en

---

[6] Estoy resumiendo aquí una conferencia titulada «Internal and External Reasons», ofrecida en Harvard, noviembre, 1978.

la obtención de la meta (en relación con sus otras metas). Esto da pie a una cuestión que tiene que ver tanto con la imaginación como con la inteligencia proposicional: *en qué consistiría efectivamente,* experiencialmente, alcanzar esa meta. Muchos seres humanos persiguen metas cuya obtención no les agradaría ni en tanta medida ni por tanto tiempo como ellos piensan. Incluso dentro de un sistema benthamita, sería posible mejorar las consideraciones en torno a la toma de decisiones racionales teniendo en cuenta la posibilidad de estimar defectuosamente el *atractivo* existencial de diversas metas. Y así se empieza a introducir un sentido en el que las mismas metas, y no únicamente los medios, pueden ser criticados como irracionales.

Una vez más, las personas *pasan por alto* metas que podrían perseguir si las tuvieran en cuenta. Aun cuando las tengan en cuenta (o si alguien se las sugiere), carecen de imaginación (¡otra vez la imaginación!) para concebir en qué consistiría la obtención de esas metas, máxime si esas metas constituyen rasgos del carácter a largo plazo, como el desarrollo de la capacidad de apreciación de la poesía. El hombre que prefiere el juego de las clavijas a la poesía puede que en efecto no sea capaz de imaginar en qué consistiría tener cierta sensibilidad hacia los matices de la auténtica poesía, y si fuera posible mejorar su inteligencia o perfeccionar su imaginación, podría convencerse de que está cometiendo un *error.*

Es significativo que la capacidad de crítica racional de las metas propias (y de las de los demás) pueda depender tanto de la imaginación como de la capacidad de aceptar enunciados verdaderos y de descreer de los falsos. Y también lo es el que una meta pueda consistir en un rasgo de la mentalidad o del carácter a largo plazo, y no en una cosa o en un acontecimiento.

Hay otras maneras de cometer errores al elegir las metas, además de estimando defectuosamente la auténtica significación experiencial de las metas o de las posibles metas alternativas. Williams señala (reviviendo una afirmación que retrotrae a Aristóteles) que muy a menudo la meta es de carácter *general* (por ejemplo, pasar un buen rato esta tarde), y el problema no es tanto encontrar *medios* para el «fin» como encontrar un patrón global de la actividad que constituya una *especificación* aceptable de la meta (por ejemplo, «ir al cine» o «permanecer en casa leyendo un libro»). Y la capacidad de pensar en especificaciones nuevas y creativas de sus metas, o sólo en especificaciones tópicas y comunes, dependerá de nuevo de la imaginación, y no sólo de la inteligencia proposicional.

El problema, tal como apunta Williams, es que aunque reemplacemos la estrecha psicología benthamita por una descripción que haga justicia a todas estas cosas, parece que aún resta cierto relativismo. El ejemplo de Williams consistía en el caso hipotético de un joven cuyo padre desea que curse la carrera militar. El padre apela a

la tradición familiar (los varones habían sido oficiales del ejército durante generaciones) y al patriotismo, pero en vano. Aun cuando el joven se represente tan vívidamente como pueda en qué consiste ser un oficial del ejército, no hay nada en esta meta que le atraiga. Esta no es precisamente *su* meta, y no por algún fallo de su inteligencia o de su imaginación.

Hasta el caso del nazi podría ser semejante. Supongamos que los nazis hubieran ganado la guerra, de modo que no podremos apelar a la derrota de Alemania como una razón decisiva para no ser nazi. Quizá algunos nazis fueron nazis simplemente porque desconocían las consecuencias reales del nazismo, el sufrimiento que acarrearía, etc. Quizá algunos nazis no hubiesen sido nazis si hubieran tenido la suficiente inteligencia e imaginación para apreciar estas consecuencias, o para apreciar vívidamente una vida alternativa, una vida pacífica. Pero sin duda muchos nazis habrían sido nazis a pesar de todo, porque no les habría importado el sufrimiento que causarían sus acciones, y porque, sin importar el grado de vividez con que se hubiesen imaginado la vida alternativa, ésta no les diría más a ellos que la vida militar lo hacía en el caso del joven del relato de Bernard Williams. No hay *en ellos* ningún fin al que apelar, ni actual ni potencial, un fin que llegarían a realizar si fueran más inteligentes e imaginativos. Incluso sin «psicología benthamita», nos enfrentamos de nuevo con el problema del relativismo moral.

Consideremos un caso menos irritante que el de los nazis. Imaginemos una sociedad de granjeros los cuales, por algún motivo, no tienen ningún interés ni en las artes, ni en la ciencia (excepto en aquellos productos que les ayudan en la labranza), ni en la religión, en resumen, en nada espiritual o cultural. (No pretendo insinuar que en realidad las sociedades campesinas sean o hayan sido así.) No necesitamos imaginar que estas gentes sean unos *rufianes*. Imaginémoslos, si nos place, como gente cooperativa, pacífica y razonablemente amable entre sí. Lo que quiero que imagine el lector es que sus *intereses* se limitan a metas tan mínimas como conseguir el alimento suficiente, conseguir un cálido refugio, y placeres tan simples como emborracharse juntos por las tardes. En resumen, imaginémosles viviendo una existencia «relativamente animal» y sin desear vivir otro tipo de existencia.

Esta gente no es *inmoral*. En su modo de vida no hay nada impermisible. Pero nuestra tendencia natural (a no ser que nos haya hechizado el relativismo ético) es decir que su modo de vida es de alguna manera despreciable. Carece totalmente de lo que Aristóteles llamaba «nobleza». Llevan una vida de puercos —quizá de puercos afables, pero, pese a todo, de puercos, y la vida de un cerdo no es vida para un hombre.

Al mismo tiempo —y aquí reside la dificultad— no nos inclina-

mos a afirmar que los hombres-cerdo sean *irracionales*. Ello puede ser resultado de nuestra prolongada aculturización en el uso benthamita de «racional» y de «irracional», pero, de cualquier modo, es nuestra disposición actual. Queremos decir que las vidas de los hombres-cerdo podrían ser mejores de lo que lo son, pero no que son *irracionales*.

No queremos decir que llevar una vida mejor o peor sea meramente una cuestión de gusto. No vemos el modo de afirmar que es *racional* elegir la vida mejor e irracional la peor. Con todo, dejar de decir tales cosas es semejante a estar diciendo que «todo es relativo». El suelo se desmenuza bajo nuestros pies.

Quizá algunas de las correcciones que Bernard Williams sugiere para la psicología benthamita nos ayuden en este caso. Supongamos que los hombres-cerdo han nacido con el *potencial* humano normal (si no fuese así sus vidas *no serían* «peor de lo que podrían ser», y nada justificaría nuestro desprecio, sino sólo, como máximo, nuestra compasión), entonces se les podría inducir a que apreciasen los aspectos artísticos, científicos y espirituales de la vida; a llevar una vida más genuinamente humana, por así decirlo. Y si alguno de ellos llegase a apreciar estos aspectos no hay duda de que *preferirían* esa vida (aun cuando podría ser menos divertida) a la que ahora viven. La gente que lleva una vida porcina siente *vergüenza* cuando llega a vivir una vida más humana (la gente que lleva una vida más humana no se *avergüenza* de haberla llevado cuando se hunde en esa vida animal). Estos hechos nos dan motivos para pensar que los hombres-cerdo están cometiendo el tipo de error, de deficiencia cognitiva, del que hablaba Williams; motivos para pensar que han pasado por alto metas alternativas y sin duda también para pensar que nunca se *representaron vívidamente* en que consistiría la realización de esas metas alternativas. En resumen, no se puede decir que hayan *escogido* peor vida, porque nunca han tenido una adecuada concepción de algo mejor.

Mientras que estas consideraciones podrían apoyar la idea de que la vida de aquellas gentes está abierta a crítica racional, no es evidente cómo aplicarlas al caso del nazi. (Se podría establecer la tautología de que quien en realidad no *escoge* la vida mejor no la ha «concebido adecuadamente», pero tal maniobra no nos serviría de ayuda, claro.) Hasta en el caso de los hombres-cerdo, si fueran hombres-cerdo ideológicamente fanáticos, y no meros hombres-cerdo, entonces nuestra indicación que apelaba a la vergüenza no se sostendría. En tal caso, podría ser que no tuviesen ningún fin, ni siquiera latente, al cual apelar.

Nuestra reluctancia a acusar a los hombres-cerdo de un defecto en la *razón* (a no ser que podamos señalar que tienen algún fin, por lo menos latente, que no logran realizar) es producto de las recientes vicisitudes de la concepción de la razón en nuestra cultura, como es

fácil establecer. Pues ni los filósofos antiguos ni los medievales veían extraño decir que si *A* es una vida *mejor* que *B*, entonces este hecho es una razón, la mejor de las posibles, para *elegir A* frente a *B*. Hemos perdido la capacidad de ver cómo la bondad de un fin puede hacer *racional* elegir tal fin.

Por supuesto, esto se explica en gran medida por el hecho de que no consideramos la «bondad» como algo objetivo. Pero ahora nos vemos enfrentados a un círculo, o más bien a dos curvas. Está el circulo moderno: la concepción instrumentalista de la racionalidad respalda la pretensión de que la bondad de un fin no hace particularmente irracional el que no se escoja, o que se escoja un fin que es manifiestamente malo, lo cual respalda la pretensión de que la bondad y la maldad no son objetivas, lo cual respalda a su vez la pretensión de que la concepción instrumentalista de la racionalidad es la única inteligible. Y está el arco tradicional: la razón es la facultad de escoger fines sobre la base de su *bondad* (en oposición a las pasiones, que intentan dictar normas sobre la base de los apetitos o «inclinaciones»), pretensión que apoya la opinión de que es *racional* elegir lo bueno, la cual respalda a su vez la pretensión de que la bondad y la maldad son objetivas. Evidentemente no podemos retroceder hacia la cosmovisión antigua o medieval, como podrían desear los conservadores; pero ¿es el círculo benthamita la única alternativa que en realidad nos queda?

8. EL IMPACTO DE LA CIENCIA EN LAS CONCEPCIONES MODERNAS DE LA RACIONALIDAD

Si la discusión que hemos examinado —una discusión que dura bastantes décadas— parece inconclusiva quizá sea porque siempre presupone cierto tipo de prioridad de la racionalidad sobre la bondad. La pregunta es siempre si hay algún sentido en el que pueda llamarse «irracional» elegir un mal fin, como si la bondad estuviera sometida a juicio y la racionalidad fuera el juez. Adoptar esta postura, especialmente cuando nuestros presupuestos con respecto a la racionalidad son, en gran parte, una colección inexaminada de mitos y prejuicios culturales, es juzgar de antemano la cuestión del *status* de los juicios de valor. Propongo invertir los términos de la comparación en lo que resta de este ensayo, no preguntando ¿cuán racional es la bondad?, sino ¿por qué es bueno ser racional? La pregunta por el valor de la racionalidad nos obligará a clarificar su naturaleza y a arrojar luz sobre los presupuestos que tendemos a admitir con respecto a la racionalidad, y puede permitirnos comprender cuál es la causa de nuestros errores al afrontar la pregunta anterior.

Recordemos que cuando Max Weber introdujo la moderna distinción entre hechos y valores, su argumento en contra de la objetividad de los juicios de valor consistía precisamente en que no es posible establecer la verdad de un juicio de valor de un modo satisfactorio para *toda posible persona racional*[1]. Ya desde el principio era la imposibilidad, o la pretendida imposibilidad de prueba racional, la que arrojaba una luz un tanto sospechosa sobre los juicios de valor. La racionalidad ha estado llevando a juicio al valor desde hace mucho tiempo. Y en este contexto racionalidad siempre significa racionalidad científica[2]; se afirma que los resultados de la ciencia positiva son

---

[1] En particular «Die Objektivität sozialwissenschaftlicher Erkenntnis», en *Archiv für Sozialwiss und Sozialpolitik*, vol. 19, 1904, pp. 24-87, y «Der Sinn der Wertfreiheit der soziologischen un ökonomischen Wissenschaften», en *Logos*, VII, 1917, pp. 49-88, y «Wissenschaft als Beruf», Vortrag, 1919. Estos tres textos se reimprimieron en *Methodology of the Social Science*, Illinois, 1949.

[2] K. O. APEL lee a Weber como yo en «The Common Presuppositions of Hermeneutics and Ethics: Types of Rationality Beyond Science and Technology», *Research of Phenomenology*, n.° IX, 1979. Apel escribe lo siguiente (p. 36): «No obstante, Max Weber también propuso una respuesta estrictamente negativa en relación con mi pregunta por tipos posibles de racionalidad que se hallen fuera de la ciencia y de la tecno-

los únicos que pueden establecerse satisfactoriamente para todas las personas racionales. Hay una razón obvia para valorar la racionalidad. Es innegable que la racionalidad científica nos ayuda a alcanzar diversas metas prácticas. Si bien pocas personas cultas subscribirían la opinión de que debemos perseverar en la ciencia únicamente por el éxito tecnológico, no hay duda de que el éxito tecnológico de la ciencia es, en el sentido más literal de la palabra, arrollador. Vivimos una serie aparentemente interminable de revoluciones tecnológicas —«la revolución industrial», «la revolución electrónica»— que constantemente nos recuerdan cuánta importancia tiene la fuerza de la ciencia en la configuración de nuestras vidas. Incluso antes de la revolución industrial, el éxito aparentemente excepcional de la física newtoniana impresionó a unas cuantas inteligencias. Por ejemplo, cuando en el siglo XVII comenzaba a discutirse la noción de «progreso», los progresistas afianzaron su posición afirmando que «Newton sabía más que Aristóteles». Nadie podía argumentar convincentemente que Shakespeare era mejor dramaturgo que cualquiera de los trágicos de la antigüedad, pero parecía innegable que el científico Newton había conseguido un innegable avance en el conocimiento en relación con el científico Aristóteles.

Pese a que los enciclopedistas y otros pensadores se apresuraron a generalizar la noción de progreso desde la ciencia hasta las instituciones políticas y la moralidad, esa generalización le ha parecido tan dudosa del siglo XX como evidente al XIX.

---

logía, ambas libres de valores. Lo que sugiero es que esta respuesta se ha convertido en un paradigma para el actual sistema ideológico de Occidente. Weber restringió el alcance de la comprensión explicativa a una comprensión "axiológicamente neutral" que centró en torno al "tipo ideal" de "comprensión propositivo-racional" de las "acciones propositivo-racionales". Ahora bien las "acciones propositivo-racionales" pueden llamarse también "acciones instrumentales"; y en aquellos casos en los que estas acciones tienen éxito, pueden ser analizadas o reconstruidas como acciones que se basan en transposiciones exitosas de las reglas del tipo "si-entonces" de la ciencia nomológica a las reglas del tipo "si-entonces" de las prescripciones tecnológicas. Por lo tanto, Max Weber restringió de este modo la función de la comprensión explicativa al intento de *captación de la racionalidad tecnológica de medios-fines* tras las acciones humanas, y es esta idea de *racionalidad instrumental* la que constituye en realidad el paradigma weberiano de racionalidad. Debe señalarse, sin embargo, que para una *comprensión propositivo-racional* en sociología no es necesario satisfacer el requisito máximo de asegurarnos que el agente logró transponer las reglas nomológicas a sus máximas tecnológicas con respecto a las relaciones medios-fines. Para comprender sus acciones a la luz de ese tipo de racionalidad instrumental basta con asegurarnos que era racional para el agente actuar como lo hizo bajo la presuposición de que *sus* propósitos y *sus* creencias con relación a los medios o estrategias eran los adecuados para alcanzar sus propósitos. De modo que la conjetura hipotética y la verificación de aquellas metas-intenciones y medios-creencias del agente, bajo cuya luz sus acciones pueden comprenderse como racionales, en el sentido de la racionalidad tecnológica medios-fines, se convierte en una función de la comprensión empírico-hermenéutica.»

Augusto Comte construyó una filosofía, el positivismo, que celebraba el éxito de la ciencia. La historia es la historia de un triunfo: comenzamos con mitos primitivos, éstos se refinaron y purificaron hasta que finalmente aparecieron las grandes religiones, éstas dieron lugar a su vez a las teorías metafísicas de Platón o Kant y, en nuestros días, la misma metafísica tiene que dejar lugar por fin a la «ciencia positiva»; evidentemente, no hay duda de quién es la heroína de esta triunfante historia: la heroína es la ciencia. Y si lo que desde el principio impresionó a unos pocos fue el pasmoso éxito intelectual de la ciencia, no hay duda de que lo que ha impresionado a la mayoría es su arrollador éxito tecnológico y material. Y nos impresiona pese a amenazar nuestras vidas.

Así pues, una razón para dudar de que los juicios de valor tengan *status* cognitivo es que no pueden «ser verificados por los métodos de la ciencia», tal y como se nos ha repetido una y otra vez. A ello también contribuye el hecho de que sea imposible obtener un acuerdo universal, ni siquiera mayoritario, en cuestiones éticas, hecho que ya hemos visto que desempeñaba un papel central en la discusión de Foucault. No parece haber respuestas universalmente satisfactorias con respecto a las cuestiones de si la homosexualidad o el aborto son correctos o no; en cambio, es creencia general que la corrección de una teoría científica puede ser demostrada con el beneplácito de todos. Algunas veces se considera que la misma racionalidad de la ciencia consiste, en parte, en el hecho de que por lo menos sus predicciones pueden ser demostradas públicamente, y que todo el mundo está dispuesto a admitir que se obtendrán tales resultados, esto es, que ocurrirán los fenómenos que la teoría predice. En este punto existe, claro, una amenaza de circularidad: si *identificamos* los procedimientos racionales como aquellos procedimientos que conducen a conclusiones admisibles para la gran mayoría, entonces el argumento de Weber de que no puede obtenerse la unanimidad de todas las personas racionales con respecto a las cuestiones éticas, aun siendo correcto, significaría que, en estas cuestiones, no puede obtenerse la unanimidad de todos aquellos que usan los métodos garantizados para obtener el asentimiento de la mayoría o de una aplastante mayoría. Es decir, el modo de determinar que los juicios de valor no pueden ser verificados con el beneplácito de todas las personas *racionales* consiste, simplemente, en observar que no pueden ser verificados con el beneplácito de la mayoría aplastante de *todas* las personas. Y esto, después de todo, no es disponer de una *prueba* para la racionalidad. La formulación de Weber sugiere que, de algún modo, primero hacemos un recuento de aquellos miembros de la población que son racionales y, después, observamos si todos ellos están dispuestos a admitir que algún juicio de valor es verdadero o no lo es. Pero, en realidad, las cosas no son así. Todo lo que muestran los ejemplos de Weber (chinos-

mandarines, etc.) es que los juicios de valor no pueden ser verificados con el beneplácito de todas las personas cultas o inteligentes (que de ninguna manera son las mismas que todas las personas *racionales*). El argumento de Weber es, de forma velada, un argumento mayoritarista: apela al hecho de que podemos obtener el acuerdo de la gente culta en relación con la «ciencia positiva», mientras que no podemos obtener tal acuerdo con respecto a los valores éticos. Es interesante contrastar esta posición con la de Aristóteles: el estagirita afirmaba que siempre hemos de intentar obtener el acuerdo con la mayoría, a pesar de que, siendo realistas, sabemos que frecuentemente no podemos alcanzarlo. Algunas veces sólo somos capaces de convencer al sabio, aunque decir esto suene elitista a los oídos de nuestros días; y, por supuesto, hemos de confiar en nuestro juicio para distinguir entre quiénes son sabios y quiénes no lo son.

No es del todo cierto que podamos obtener un acuerdo aplastante en lo que respecta a la verdad de una teoría científica aceptada. En realidad, la mayoría de la gente lamentablemente ignora muchas teorías científicas, en especial de las ciencias exactas, cuya comprensión requiere tanta matemática que la mayoría ni siquiera son capaces de entenderlas. Y, claro está, aunque este punto se admite, no parece afectar a la mayoría de las personas, pues, de acuerdo con el aguado operacionalismo en que parece haberse convertido la filosofía con la que trabajan muchos científicos, el contenido de una teoría científica consiste en consecuencias contrastables que pueden expresarse mediante enunciados de la forma siguiente: *si realizamos tales y tales acciones, obtendremos tales y tales resultados observables*. La verdad de estos enunciados puede demostrarse, cuando son verdaderos, repitiendo el experimento apropiado con la suficiente frecuencia. Esta consideración tropieza con muchas dificultades: los experimentos son mucho más difíciles de diseñar, realizar y evaluar de lo que pueden creer los legos. Pero sin duda, como cuestión de hecho, ha sido posible alcanzar un amplio acuerdo sobre la adecuación experimental de ciertas teorías de las ciencias exactas. Que los legos acepten estas teorías puede deberse a su deferencia con los expertos, pero al menos los expertos parecen estar de acuerdo.

El instrumentalismo no constituye por sí mismo una concepción de la racionalidad intelectualmente sostenible, desde luego. Nadie duda que los resultados científicos tienen un enorme valor práctico, pero, como ya hemos dicho, ninguna persona culta sostiene que el único valor de la ciencia lo constituyan sus aplicaciones. Y aun cuando se la valorase únicamente por sus aplicaciones, ¿por qué ha de valorarse la *racionalidad* por sus aplicaciones? Sin duda, es de gran valor tener un instrumento que nos ayude a seleccionar medios eficientes para la obtención de nuestros diversos fines; pero también es valioso saber qué fines hemos de escoger. No es sorprendente que la verdad

de los juicios de valor no pueda ser «racionalmente demostrada» si la «verificación racional» está limitada, por definición, al establecimiento de las conexiones medios-fines. Pero, en primer lugar, ¿por qué hemos de mantener una concepción de la racionalidad tan estrecha? [3].

El mayoritarismo también es insostenible. Por supuesto que es agradable obtener el acuerdo con respecto a lo que uno considera que es verdadero. Y también es siempre agradable evitar el conflicto con los colegas. Mas la gente ha vivido durante siglos con el incómodo conocimiento de que en ciertos asuntos uno tiene que confiar en su propio juicio, aun cuando éste difiera del juicio de la mayoría. Muchos se han enorgullecido de confiar en su propio juicio, aunque éste haya sido diferente del de la mayoría. La idea de que en algunas cuestiones, entre las cuales se hallan las cuestiones éticas, han de superarse consideraciones tan complejas e imprecisas que no podemos tener la esperanza de contar con pruebas o definiciones científicas, sino que tenemos que contar con la perspicacia y el buen juicio, no es ninguna idea nueva. Y es plausible que una de las manifestaciones más altas de la racionalidad sea la capacidad de juzgar correctamente en aquellos casos en los que no se puede esperar «probar» cosas. Parece realmente extraño que el hecho de que ciertas cosas no pueden probarse satisfactoriamente para todos se haya convertido en un argumento a favor de la irracionalidad de las creencias sobre estas cosas [4].

Pese a la debilidad intelectual de estas concepciones, tanto el instrumentalismo como el mayoritarismo ejercen un poderoso influjo sobre la mentalidad contemporánea. A ésta le gustan los éxitos demos-

---

[3] Atribuyéndole a Weber esta concepción estrecha de la racionalidad, Apel escribe lo siguiente (*loc. cit.* p. 37): «Este punto de la metodología weberiana de la "comprensión" estaba en perfecta concordancia con su filosofía de la historia (más o menos explícita). Pues, en el contexto de su reconstrucción de la historia de la civilización occidental, partió de la hipótesis heurística de que al menos esta parte de la historia podía concebirse como un continuo progreso de "racionalización" y, al mismo tiempo, un proceso de desilusionamiento, o, como le gustaba decir, de "desencantamiento" [*Entzaüberung*]. Por "racionalización" entendía el progreso en la puesta en vigor de la racionalidad medios-fines en todos los sectores socio-culturales, especialmente en la esfera de administración burocrática y económica, bajo la constante influencia del progreso científico y económico. Por otra parte, por proceso de "desilusionamiento" o "desencantamiento" Weber entendía la disolución del orden de valores o de la cosmovisión religiosa o filosófica comúnmente aceptada. El estaba dispuesto a extraer las consecuencias prácticas de este proceso para su cosmovisión personal, hasta el punto de sugerir que un pensador riguroso y sincero tenía que aceptar la idea siguiente: el progreso humano, en el sentido de "racionalización", se complementa con el abandono de la idea de imposición racional de normas y valores últimos, en favor de la adopción del recurso a decisiones de conciencia pre-racionales y últimas, de cara a un pluralismo, o, como dijo Weber, un "politeísmo" de normas o valores últimos.»

[4] En realidad, ya vimos en el capítulo 5 que la teoría de la aceptabilidad racional por consenso se autorrefuta.

trables, y se siente incómoda ante las mismas nociones de buen juicio y sabiduría. No soy sociólogo y no intentaré investigar la cuestión de por qué la sociedad industrial, tanto en sus versiones capitalistas como socialistas, ha de estar tan ligada a los temas del éxito instrumental y del consentimiento mayoritario. Pero el hecho sociológico tiene que ver indudablemente con la importancia siempre en aumento de la concepción que hace equivalentes racionalidad y racionalidad científica, y de la concepción que basa la racionalidad científica en la demostración de conexiones instrumentales con el convencimiento (potencial) de una aplastante mayoría.

Si bien la concepción de la racionalidad que acabamos de describir, a saber, la idea de que la racionalidad consiste en métodos (cuya naturaleza queda bastante en el aire) que dan por resultado el descubrimiento de conexiones efectivas entre medios y fines y en el establecimiento «público» de estas conexiones, tal como está, es intelectualmente insostenible, no han faltado intentos filosóficos por hacerla respetable. Uno de estos intentos nace del más antiguo empirismo de Locke, Berkeley y Hume. Por la época de Mill, este empirismo se había solidificado en lo que los filósofos llaman fenomenalismo: la doctrina que afirma que, en realidad, sólo podemos referirnos a *sensaciones*. Según esta concepción, hasta los objetos cotidianos, mesas y sillas, son de hecho solamente conjuntos de regularidades objetivas de sensaciones humanas actuales y posibles. Como expuso Mill, los objetos físicos son «posibilidades permanentes de sensación». Otro modo de exponer la misma idea es decir que todo discurso que parece referirse al mundo físico es únicamente un discurso referente a sensaciones, si bien muy derivado.

Este punto de vista tenía la virtud, a los ojos de sus adalides, de que les permitía decir con claridad cuál era el contenido no sólo de la ciencia, sino también de cualquier discurso cognitivamente significativo. Cualquier teoría científica es sólo un medio «económico» de establecer una cantidad de hechos de la forma: *si realiza tales y tales acciones, entonces obtendrá tales y tales experiencias.* El partidario de esta concepción no tiene por qué defender la insostenible afirmación de que los científicos sólo están interesados en las aplicaciones, o en la obtención de metas prácticas, y no en el conocimiento por sí mismo. El fenomenalismo no tiene por qué negar nuestro deseo de conocer la naturaleza de los agujeros negros, de saber si hubo o no *Big-Bang,* o de saber cuál fue el auténtico origen del *homo sapiens.* Queremos saber todas estas cosas, y no sólo porque su conocimiento nos permita construir mejores máquinas. Pero conocer estas cosas es, aunque para demostrarlo se requiera un refinado análisis filosófico, conocer una gran cantidad de hechos de la forma: *si realiza tales y tales acciones, entonces tendrá tales y tales experiencias.* Cualquiera que sea el motivo por el que nos interesen, todos los hechos *son en*

última instancia instrumentales. Al mismo tiempo, no parece que haya manera de descifrar la afirmación de que llamar algo «bueno» es llevar a cabo una predicción de la forma: *si realiza tales y tales acciones, entonces tendrá tales y tales experiencias.* De modo que, según esta concepción, los enunciados con respecto a la bondad o la maldad de algo no tienen significado cognitivo; dicho con palabras de los empiristas lógicos del siglo veinte, tales enunciados son «puramente emotivos». No obstante, el fenomenalismo fracasó en dos puntos. En primer lugar, la afirmación de que todos los enunciados sobre objetos físicos son traducibles a enunciados sobre sensaciones actuales y posibles parece ser, de hecho, falsa. Una meticulosa investigación lógica, comenzando por la obra de Carnap y del Círculo de Viena en los años 30, convenció a los fenomenalistas de que esa afirmación carecía de fundamento. Las teorías científicas, entendidas como todos integrados, nos inducen a esperar que tendremos ciertas experiencias si llevamos a cabo ciertas acciones; pero la idea de que los enunciados de la ciencia son traducibles uno a uno a enunciados sobre las experiencias que tendremos si llevamos a cabo ciertas acciones, ya ha sido abandonada como un inaceptable tipo de reduccionismo. En segundo lugar, las sensaciones son necesariamente objetos privados. Aunque en la práctica sí somos capaces de decidir, mediante una simple pregunta, si alguien tuvo o no una sensación, si alguien plantea la cuestión «¿Cómo sabe usted que tal persona asocia a sus descripciones las mismas sensaciones que usted?», nos encontramos de inmediato ante un problema epistemológico. Si el contenido de la ciencia consiste en predicciones con respecto a qué sensaciones tendrá un ser racional si realiza tales acciones, entonces para saber lo que ese mismo contenido significa, si nos encontramos ante un extraterrestre, habríamos de ser capaces de decir si tiene o no las mismas sensaciones que nosotros, etc. Por esta razón filósofos como Rudolph Carnap y Sir Karl Popper insistieron en que las predicciones observacionales de la ciencia deben establecerse de esta forma: si alguien realiza tales y tales acciones, entonces *tendrán lugar tales y tales acontecimeintos públicamente observables,* donde tanto las acciones como los acontecimientos observables esperados deben describirse en términos de objetos «públicos», por ejemplo, lecturas de contador, y no en términos de objetos privados tales como las sensaciones.

En suma, el antiguo empirismo, o el fenomenalismo, parecía proporcionarnos un pulcro criterio de significación cognitiva: un enunciado es cognitivamente significativo si es traducible a un enunciado acerca de sensaciones. Pero resulta que o bien la noción de «traducción» es desesperadamente vaga, o bien los propios enunciados de la ciencia dejaban de satisfacer ese criterio de significación cognitiva. El problema de trazar una línea divisoria entre los enunciados fácticos y los juicios de valor sobre la base de que sólo los primeros «traduci-

bles» a enunciados acerca de sensaciones reside en que la pretendida traducibilidad de la primera clase de enunciados no ha sido demostrada y, aparentemente, no puede serlo. El reduccionismo empirista trazó una línea divisoria entre lo fáctico y lo evaluativo, pero al precio de ofrecer una descripción completamente deformada de lo fáctico.

No obstante, nuestro propósito original fue considerar qué respuestas se habían dado a la pregunta «¿Por qué es bueno ser racional?». La primera respuesta que examinamos y rechazamos por ser demasiado estrecha afirmaba que la racionalidad nos permite descubrir conexiones fiables entre medios y fines. El fenomenalismo entró en escena porque, de ser verdadero, disolvería el conflicto entre el interés en una teoría científica por sus consecuencias instrumentales y el interés por aprender lo que esta teoría nos cuenta con respecto a los procesos naturales. El conflicto entre intereses instrumentales e intereses puramente teoréticos podría ser hasta cierto punto artificial. Perduraría algún tipo de diferencia entre estos intereses, desde luego; pero hasta los intereses puramente teoréticos serían intereses por hechos cuya naturaleza, en última instancia, se habría revelado instrumental. *Todo* conocimiento que merezca este nombre se habría mostrado como un conocimiento de las conexiones medios-fines. Sólo llamamos «prácticos» a nuestros intereses cuando estamos interesados en la conexión medios-fines porque esperamos explotarla de cara a la obtención de alguna meta, y los llamamos «teóricos» cuando nos interesa conocer la conexión medios-fines por pura curiosidad. Como hemos señalado, este intento de reducir cada enunciado científico a un enunciado de la forma *si realiza la acción A, entonces obtendrá el resultado B,* ha fracasado.

La propuesta de Carnap y Popper, que afirmaba que los enunciados observacionales de la ciencia se expresan en un lenguaje de objetos físicos y no en un lenguaje de sensaciones, es obviamente correcta cuando se considera como una generalización con respecto a la práctica de los científicos. No obstante, cuando se erige en un absoluto epistemológico, su importancia llega a ser aun más decisiva. Por un lado, si ningún enunciado observacional está autorizado a referirse a sensaciones, entonces se rechaza la introspección como medio de observación científica. Pese a que muchos psicológos estarían de acuerdo en que debe ser rechazada, lo cierto es que otros, sin amedrentarse ante tal dogma filosófico y psicológico, han seguido realizando experimentos que entrañan, al menos en parte, cierta confianza en los informes introspectivos. De hecho, Carnap no habría sido tan dogmático en esta prohibición como lo fueron algunos psicólogos behavioristas; él habría permitido el uso de informes fenomenalistas, con tal de que no se analizasen como informes de observación sino como datos conductuales, siendo la «conducta» la elaboración de esos mismos informes. Pero no está del todo claro qué significa analizar la

aceptación de un informe introspectivo como una «inferencia a partir de la conducta verbal». Incluso en el caso de informes que no versen sobre sensaciones sino sobre objetos físicos, por ejemplo, «Hay una mesa ante mí», tampoco aceptamos normalmente el informe a menos que dispongamos de una teoría de acuerdo con la cual la persona se hallaba en condiciones de observar el hecho del que informa. En este sentido, es parte de nuestra demanda global de coherencia en nuestra imagen del mundo el que las observaciones sean teóricamente explicables; si alguien pretendiese haber observado mediante la *clarividencia* que había una mesa en cierto lugar, no aceptaríamos su «informe observacional», ya que no concuerda con nuestro corpus teórico global. En este sentido, *cada* informe observacional incorpora un componente que podríamos describir como «inferencial». Por otra parte, cuando un médico, por ejemplo, acepta el informe de un paciente que afirma sentir dolor, es difícil saber cuál es la «teoría científica» sobre cuya base el médico infiere, a partir del informe verbal del paciente, que éste siente dolor; si la suposición general de que las personas están en condiciones óptimas para decir si les duele algo o no cuenta como teoría, entonces también contará como teoría la suposición general de que las personas están en condiciones óptimas para afirmar si hay una mesa ante ellas o no la hay; pero es difícil apreciar alguna diferencia metodológica fundamental entre aceptar el informe de alguien que diga que hay una mesa ante él y aceptar el informe de alguien que diga que le duele algo.

Popper y Carnap replicarían que la diferencia metodológica estriba en que el primer enunciado es comprobable públicamente, mientras que el segundo no lo es; pero ambos exageran la medida en la que los enunciados observacionales son públicamente observables. Gran parte de estos informes se evalúan con la ayuda de instrumentos cuyo uso requiere bastante preparación. (Es sabido que aprender a «observar» a través de un microscopio de alta resolución requiere un buen grado de preparación especializada y de pericia, y que no todo el mundo es capaz de adquirirla.) La aceptación de este dogma epistemológico hace que el que las creencias racionales sean susceptibles de comprobación *pública* sea parte de la definición de racionalidad. Y ello es conveniente, pues hace innecesario proporcionar un argumento en favor de esta propuesta. Quizá el argumento consista en el fondo en afirmar que lo que no es públicamente comprobable puede convertirse en materia de desacuerdo, y por otra parte, que siempre que hay un desacuerdo irresoluble no hay corrección o incorrección. Pero esto sería asumir lo que he llamado mayoritarismo, es decir, la idea de que en la propia noción de racionalidad está implícito que lo que es racionalmente verificable es verificable con el beneplácito de la aplastante mayoría.

En mi opinión, en la última etapa del empirismo lógico se hace

evidente el hecho de que este movimiento fue, fundamentalmente, una expresión sofisticada y generalizada de una tendencia cultural hacia el instrumentalismo y el mayoritarismo. Pese a que los empiristas lógicos habían abandonado el fenomenalismo en fecha temprana, en 1936, durante los veinte años siguientes, es decir, hasta que el movimiento comenzó a quebrarse y a desaparecer como tendencia filosófica reconocible, los filósofos de la ciencia empiristas-lógicos se aficionaron a hablar del «objetivo de la ciencia» y a identificarlo con la *predicción* (con algunos añadidos que discutiremos en su momento). Desde sus orígenes en los escritos de Augusto Comte, la idea de que el objetivo de la ciencia es la predicción fue la idea fundamental del positivismo. Como veíamos, mientras el fenomenalismo estuvo en boga, esta idea tuvo cierta base filosófica seria, ya que entonces se podía argumentar que todos los enunciados cognitivamente significativos eran predicciones disfrazadas, o conjuntos infinitos de predicciones disfrazadas. La reaparición de esta doctrina después de la desaparición del fenomenalismo es como la aparición de las «escenas primitivas» en las asociaciones de un paciente en terapia, después de que las «defensas» se han desmantelado. Decir que el objetivo de la ciencia es el éxito en la predicción (o el éxito en la predicción más algo descrito como «simplicidad») parece algo peligrosamente cercano a decir que la ciencia se persigue sólo por metas prácticas; y ningún filósofo ha deseado mantener esta posición. En realidad, los filósofos que defendieron una concepción puramente instrumental de la ciencia no lo hicieron porque fuesen devotos de lo práctico u hombres de mentalidad estrecha que no apreciasen la belleza del conocimiento científico abstracto, sino porque sintieron que identificando lo que es «cognitivamente significativo» con lo que tiene valor para la elaboración de predicciones, podrían acabar de una vez con toda forma de oscurantismo y metafísica. Para estos filósofos «metafísica» era simplemente otro nombre para denominar los diferentes tipos de especulación trascendental; lo que les espantaba eran las especulaciones religiosas y «metafísicas» (en el sentido que ellos daban a «metafísica»).

Lo que estoy sugiriendo es que, dado el elevado prestigio que la ciencia tiene en nuestra cultura, y dado el ocaso de la religión, de la ética absoluta y de la metafísica transcendental, era de esperar la aparición en nuestra cultura de una tendencia filosófica hipnotizada por el éxito de la ciencia hasta tal punto que no podía concebir la posibilidad del conocimiento y de la razón fuera de lo que nos complace llamar «ciencias». Estoy insinuando que el elevado prestigio que la ciencia tiene en *nuestra cultura* se debe en gran medida a su enorme éxito instrumental, y al hecho de que la ciencia parece estar libre de los debates interminables e irresolubles que hallamos en la religión, la ética y la metafísica.

No obstante, ya que los filósofos profesionales que racionalizaron la tendencia instrumentalista en nuestra cultura no fueron personas con una mentalidad vulgar ni personas puramente prácticas, no es sorprendente que se sintieran obligados a ampliar un poco la descripción del «objetivo de la ciencia» con el fin de dar cabida más explícitamente a otros objetivos además del éxito en la predicción. Y de esta forma encontramos otros objetivos que los autores empiristas-lógicos enumeraron en los años cuarenta y cincuenta: el descubrimiento de leyes, la retrodicción (esto es, la predicción de eventos pasados, en oposición a la de eventos futuros) y el descubrimiento de «explanaciones», por lo cual entendían sencillamente la deducción de predicciones y retrodicciones a partir de leyes.

Lo que sucedió en este punto es interesante. Con vistas a hacer explícito que a la ciencia le interesa el descubrimiento de leyes de la naturaleza por ser tales, y no meramente por las predicciones a las que éstas conducen, esos autores reemplazan la fórmula «el objetivo de la ciencia es el éxito en la predicción» por una *lista*. De hecho, la lista no tiene límites fijos: las leyes de la naturaleza resulta que incluyen no sólo leyes de la naturaleza en sentido estricto, es decir, enunciados imposibles de falsar físicamente, sino también las presuntas «leyes» de la teoría de la evolución, que son en realidad descripciones de ciertas tendencias que podrían dejar de mantenerse en algún momento debido a la acción de la vida inteligente, e incluso enunciados referentes a disposiciones puramente contingentes de los grupos de organismos y hasta de los organismos individuales. Por supuesto, es perfectamente correcto afirmar que los científicos intentan descubrir «leyes de la naturaleza», incluyendo generalizaciones físicamente contingentes que se mantienen durante largos períodos de tiempo y que tienen una amplia significación explicativa, como aquellas en las que se basa la teoría de la evolución y la ciencia económica, y que se esfuerzan en descubrir verdades significativas con respecto a las disposiciones de los grupos de organismos y de los organismos individuales y en organizar todo esto en una estructura deductiva (e inductiva). Pero ¿por qué esta lista tan particular?

El motivo es que se pensaba que esta lista era lo suficientemente amplia como para poder incluir todos los géneros de verdades que los científicos pretenden descubrir, en la ciencia física, por descontado, y lo suficientemente estrecha como para no incluir material objetable («cognitivamente insignificante»). La búsqueda de un criterio de significación cognitiva, como «Una oración es plenamente significativa si y sólo si es posible verificarla o falsarla», ha sido reemplazada por la de una lista de tipos de enunciados, de forma que un enunciado es admisible si pertenece a uno de estos tipos, y de lo contrario ha de ser rechazado. Pero ¿cómo puede un filósofo realizar esta maniobra con verosimilitud? Aun cuando sea cierto que todos los enuncia-

dos de las disciplinas que denominamos «ciencias» sean de estos tipos —y no está nada claro que éste sea el caso—, ¿es en realidad la explicación histórica una mera subsunción bajo «leyes» de una serie de retrodicciones?, ¿acaso se sigue de esto que el objetivo de la propia razón es la verificación de estos tipos de enunciados, y no únicamente el objetivo de las aplicaciones especiales de la razón a las que llamamos ciencias? La respuesta es que estos filósofos evidentemente no dudaban de que la «ciencia» agotaba la razón. ¿Pero por qué no lo dudaban? Pues porque para ellos la oposición no se establecía entre la ciencia, en el sentido de un conocimiento que procede esencialmente mediante los métodos de las ciencias empíricas y matemáticas, y la razón informal, que procede mediante métodos que podrían adaptarse a intereses diferentes de los de esas ciencias, pero que no por ello es menos capaz de poseer criterios legítimos. La oposición se establecía, más bien, entre el conocimiento que procede mediante los métodos de las ciencias y el pseudoconocimiento que pretende proceder mediante revelación o mediante algún tipo de rara facultad trascendental. La razón tenía que ser coextensiva con la ciencia, pues *¿qué otra cosa podría ser?* Sin embargo, esta pretensión colocó a esos filósofos ante singulares aprietos. Puesto que no deseaban negar el conocimiento histórico, se comprometieron con la posición que hacía de la historia una ciencia, e incluso con la posición que afirmaba que la pretensión real del historiador es subsumir bajo *leyes* enunciados singulares referentes al pasado —una afirmación sobre la historia que a primera vista parece falsa.

Quizá no sea tan sorprendente que la tendencia lógico-empirista comenzara a desintegrarse alrededor de 1950. Hemos estado examinando esta tendencia únicamente bajo la óptica de una pregunta; los empiristas lógicos tenían gran número de intereses filosóficos diferentes y realizaron muchas contribuciones valiosas. Sin embargo, bajo la óptica de la pregunta que estábamos planteando, es decir, «¿Por qué es buena la racionalidad?», el movimiento del empirismo lógico representó una defensa filosófica razonada de la opinión según la cual la respuesta, la única respuesta a la pregunta, es que la racionalidad es buena para el descubrimiento de las conexiones medios/fines. La doctrina fenomenalista proporcionó a los empiristas-lógicos una interesante defensa filosófica de esta pretensión. Cuando se abandonó el fenomenalismo y cuando la *defensa* filosófica de la pretensión fue reemplazada por la pura pretensión y, en mayor medida, cuando ésta se convirtió en algo más «razonable», permitiendo las excepciones, desapareció todo el poder incisivo de este movimiento. La posición que mantiene que los objetivos de la razón son el descubrimiento de predicciones, retrodicciones, leyes de la naturaleza y la sistematización de todo este material, y que éstos son *todos* los objetivos de la razón, tropieza con el problema de que, sencillamente, no hay razón

alguna para creerlo (y no pretendo afirmar que haya razones para creer que es falso); si el concepto de ley de la naturaleza incluye el descubrimiento de enunciados disposicionales acerca de organismos individuales, y la noción de disposición es tan amplia (o tan vaga) que el enunciado de que cierto científico envidia la reputación de su colega cuenta como un enunciado de «disposición», y si el enunciado de que tal científico se burló porque estaba celoso de su colega es una «subsunción de un acontecimiento particular bajo una ley», entonces puede ocurrir que todo lo que uno diga se interprete o como una formulación de leyes generales o como una subsunción de descripciones bajo leyes generales. Tal vez, hasta decir que alguien es *moralmente bueno* pueda analizarse como una adscripción de una «disposición» a ese alguien. No, el problema de intentar especificar los objetivos de la investigación cognitiva en general por medio de una lista de ese tipo es que esa lista ha de ser *construida:* si los términos de la lista se construyen de un modo más o menos literal, entonces los tipos de enunciados pertenecientes a la lista ni siquiera incluirían todas las clases de enunciados que a los científicos les interesa descubrir, y menos aún si «científico» incluye *historiador, psiquiatra y sociólogo.* Si los términos de la lista se construyen con tanta indulgencia que no hay dificultad alguna para construir los enunciados elaborados por los historiadores (y los enunciados descriptivos en el lenguaje de la psicología cotidiana) como enunciados pertenecientes a los tipos incluidos en la lista, ésta se vuelve inútil. En cualquier caso, en ausencia de alguna explicación epistemológica de por qué los enunciados de *tales tipos,* y sólo éstos, son susceptibles de verificación racional, tal lista sería sólo una mera hipótesis acerca de los límites de la investigación racional. Una mera hipótesis, en forma de lista o en cualquier otra forma, no podría tener la fuerza de exclusión que los empiristas lógicos querían que tuviese el «criterio de significación cognitiva».

EL FETICHISMO DEL «METODO»

Ya que la respuesta a la pregunta «¿Por qué es bueno ser racional?» no puede ser sencillamente que la racionalidad nos permite obtener metas prácticas, ni tampoco que la racionalidad nos permite descubrir las conexiones medios/fines, podemos examinar otra posible respuesta que ha tenido un considerable atractivo en diferentes épocas. Muchos filósofos de la ciencia han creído que la ciencia procede siguiendo un *método* distintivo: existe un método con la propiedad de que su uso nos permite descubrir fehacientemente verdades y ningún otro método tiene esta posibilidad real. Y si lo que *explica* el extraordinario éxito de la ciencia y la persistencia de la controversia en otros campos, es que la ciencia, y sólo ella, ha empleado consistentemente este método, entonces tal vez *debería* identificarse la racionali-

dad, en la medida en que exista algo de este tipo, con la posesión y el empleo de ese *método*. La respuesta a la pregunta «¿Por qué es bueno ser racional?» sería entonces que ser racional es bueno porque siendo racionales podemos descubrir verdades (sea cual sea el tipo de verdades que nos interesen), mientras que no siéndolo no tenemos ninguna probabilidad efectiva de descubrirlas, salvo por azar. Esta concepción, como la concepción instrumentalista, tuvo una historia filosófica de auge, estancamiento y ocaso. Desde la publicación de la *Lógica* de Mill en los años 1840 hasta la de *Logical Fundations of Probability* de Carnap, influyentes filósofos de la ciencia siguieron creyendo que en la ciencia empírica subyace algo como un método formal *(«lógica inductiva»)* y que un constante esfuerzo podría desembocar en una exposición explícita de este método, en una formalización de la lógica inductiva comparable a la formalización de la lógica deductiva alcanzada a partir de la obra de Frege en 1879. Si tal método se *hubiera* descubierto, entonces, y aun cuando esto no *probaría* que el método agotase la racionalidad, la carga de la prueba hubiera correspondido a aquellos que pretendían que había verdades justificables mediante algún otro método, o cuya aceptabilidad racional podría mostrarse mediante algún otro método.

De acuerdo con la escuela estadística más influyente, la llamada escuela «bayesiana», el carácter general de este método inductivo que los filósofos estaban intentando formalizar es el siguiente: supongamos o pretendamos que el lenguaje de la ciencia ha sido formalizado, y que los científicos disponen de cierto número de observaciones fiables, expresables mediante «oraciones de observación» en este lenguaje formalizado. Supongamos también que las diversas hipótesis bajo consideración se expresan mediante fórmulas de este lenguaje. El problema de la lógica inductiva se considera equivalente al problema de definir una «función de confirmación», esto es, una función probabilística que determine la probabilidad de cada una de las hipótesis en relación a la evidencia observacional, o, en otra terminología, el «grado de corroboración» que la evidencia presta a cada una de las hipótesis alternativas. Por lo general, se supone que se sabe la probabilidad de que ocurra un determinado suceso si cada una de las hipótesis alternativas fuera verdadera; esta probabilidad se llama «probabilidad *a posteriori*», es decir, la probabilidad de la evidencia *dada* la hipótesis. Lo que deseamos calcular es la llamada «probabilidad inversa», esto es, la probabilidad de la hipótesis *dada* la evidencia. El teorema de Bayes pone esta «probabilidad inversa» en función de las probabilidades *a posteriori* y de ciertas otras, las denominadas «probabilidades *a priori*» de hipótesis alternativas, es decir, las probabilidades o «grados subjetivos de certidumbre» que los científicos asignan a esas hipótesis alternativas antes de examinar la evidencia observacional.

Las «probabilidades *a posteriori*» son realmente fáciles de calcu-

lar en dos casos muy comunes: son fáciles de calcular cuando (a) la hipótesis realmente implica la evidencia (en este caso, la probabilidad «*a posteriori*» de la evidencia dada la hipótesis es 1); o cuando (b) la propia hipótesis es una hipótesis estadística o estocástica, cuyo contenido incluye que la evidencia particular obtenida debe ocurrir con cierta probabilidad *r*. La dificultad en la aplicación del teorema de Bayes —una dificultad tan seria que tanto filósofos como estadísticos están profundamente divididos con respecto a la importancia y a la utilidad del teorema de Bayes en el caso de la confirmación de teorías— es la necesidad de una métrica de la probabilidad *a priori*, un conjunto de «grados subjetivos de certidumbre», en terminología de De Finetti y Savage.

Limitémonos por el momento a hipótesis para las que pueden calcularse efectivamente las «probabilidades *a posteriori*». Para hipótesis de este tipo el método que se acaba de describir es, en realidad, un método puramente formal; esto es, podríamos programar una computadora para que calculase los grados de corroboración de las diversas hipótesis, dados los «*inputs*» apropiados. Pero los «*inputs*» tendrían que incluir no sólo las «probabilidades *a posteriori*» computables, sino también la métrica de la probabilidad *a priori* en el contexto dado. Si concebimos ésta como una representación de las creencias anteriores de los científicos acerca del mundo, como sugiere el término «función de probabilidad subjetiva», entonces da la impresión de que uno de los *inputs* del propio método es cierto conjunto de creencias fácticas substantivas (o grados de creencia) acerca del mundo. Así consideran hoy los filósofos de la ciencia el problema; se está generalizando la creencia en que no es posible trazar una línea divisoria entre el *contenido* y el *método* de la ciencia, en que el método de la ciencia cambia tan constantemente como lo hace su contenido. El teorema de Bayes, si capta realmente la lógica de la teoría de la confirmación, proporciona un modo de formalizar esta dependencia del método de la ciencia con respecto al contenido de la ciencia mediante la necesidad de una función de probabilidad *a priori*.

Para exponer la cuestión de modo algo más abstracto, podríamos decir que el fetichista del «método» supone que la racionalidad es *inseparable*. Pero el teorema de Bayes indica que éste no es el caso; que podemos separar la racionalidad incluso en el área de la ciencia, y hasta en el área especial en la que tratamos con teorías para las que son computables las probabilidades *a posteriori*, en dos partes: una parte *formal*, que puede ser esquematizada matemáticamente y programada en una computadora, y una parte *informal*, que no puede ser tan esquematizada y que depende de las creencias cambiantes de los científicos. Ahora bien, sería agradable, por no querer decir más, que la parte formal de la racionalidad bastara para garantizar buenos resul-

tados. Si fuese posible afirmar que en el caso de que los científicos efectúen sus observaciones cuidadosamente, reúnan las suficientes observaciones, y calculen los grados de corroboración de acuerdo con el teorema de Bayes, entonces llegarán a estar de acuerdo, a pesar de que al principio no lo estuviesen debido a la diferencia en sus grados subjetivos de certidumbre, en tal caso todo sería perfecto. Pero esta imagen feliz se equivoca en dos cosas.

La primera es que aun cuando podamos mostrar que a la larga la «función de probabilidad *a priori*» se elimina, es decir, que científicos con diferentes funciones de probabilidad *a priori* acabarían poniéndose de acuerdo siempre que continuaran reuniendo más evidencia y usando el teorema de Bayes, aun así sería necesario que esta convergencia fuese razonablemente rápida. Si científicos con diferentes funciones de probabilidad *a priori* no se pondrán de acuerdo hasta que el fenómeno a predecir haya tenido lugar, o hasta que hayan pasado millones de años, entonces, a corto plazo, el hecho de que haya alguna garantía matemática de una eventual convergencia no sirve de nada; el problema de las justificaciones a largo plazo es que el plazo puede ser demasiado largo. Con las célebres palabras de John Maynard Keynes, «a la larga todos estaremos muertos». El segundo error es que, de hecho, sucede que las diferencias en la función de probabilidad *a priori* pueden conducir a acerbas diferencias en los grados efectivos de corroboración asignados a las teorías, y que estas diferencias pueden ser equivalentes a lo que normalmente consideraríamos como crasas irracionalidades.

Para exponer este último punto de otra forma, diremos que un científico sólo asignará grados de corroboración a hipótesis que parezcan «razonables» si comienza con una función de probabilidad *a priori* «razonable». Si una persona tan sólo obedece a la parte formal de la racionalidad siendo lógicamente consistente y asignando grados de corroboración en conformidad con el teorema de Bayes, pero si su función de probabilidad *a priori* es extremadamente «irrazonable», entonces sus dictámenes sobre el grado en que la evidencia apoya a las diversas hipótesis serán insensatos e «irracionales» (tal como los científicos y la gente normal, en efecto, los juzga). La racionalidad formal, el compromiso con la parte formal del método científico, no garantiza la racionalidad real y efectiva.

De hecho, esto es cierto hasta el punto de rayar en el escándalo. Arthur Burks ha mostrado que existen hasta «funciones de probabilidad *a priori* contrainductivas». Esto es, exista cierta métrica de «probabilidad *a priori*», lógicamente posible, tal que si un científico tuviese esa medida, cuanta más evidencia acumulase una hipótesis (usando el término «más evidencia» sobre la base de nuestros juicios inductivos normales) el científico asignaría una importancia cada vez menor a esa hipótesis durante muchísimo tiempo.

Una solución para esta dificultad podría consistir en complementar la presente descripción formal del método científico mediante un conjunto adicional de reglas formales que determinarían qué *a prioris* son razonables (en lo sucesivo, me referiré a las funciones de probabilidad *a priori* simplemente como «*a priori*», en conformidad con el uso estadístico). Pero no parece haber ninguna razón de peso para creer que exista un conjunto de reglas que pueda distinguir entre *a prioris* razonables e irrazonables, o que éste pueda ser más simple que una descripción completa de toda la psicología de un ser humano idealmente racional. Parece haberse evaporado la esperanza en un método formal susceptible de ser aislado de los auténticos juicios humanos acerca del contenido de la ciencia (esto es, acerca de la naturaleza del mundo) y que se halle libre de los valores humanos. E incluso ampliando la noción de método, de modo que una formalización de la psicología de un científico humano idealmente racional contase como «método», no hay razón para pensar que el «método», en *ese* sentido, fuese independiente de juicios estéticos, éticos, etc. Después de todo, la única razón para creer que el método científico no era aplicable a las creencias acerca de cuestiones éticas, estéticas, etc., ni las presuponía, era la creencia de que aquél constituía un *método formal*.

Mi discusión ha dependido de dar por sentada la corrección de una particular aproximación a la formalización del método científico, el enfoque denominado «bayesiano». Pero surgen problemas similares en cada una de las restantes tentativas ensayadas. Aun cuando se intente aislar alguna pequeña parte del método científico que no sea tan «potente» como la confirmación de teorías, lo cual estaría mucho más en línea con lo que Bacon entendía por «inducción», esto es, aun cuando se intente aislar un método para confirmar simples generalizaciones y «proyectar» la verdad de la generalización, surgirán problemas similares. Nelson Goodman [5] ha mostrado que ninguna regla *puramente formal* para la proyección inductiva puede siquiera estar libre de inconsistencias: antes de que se pueda esperar que una regla formal rinda por lo menos resultados consistentes, los predicados del lenguaje deben ser segregados *de antemano* en aquellos que se desea considerar como «proyectables» y aquellos que se tratará como «no-proyectables». El hecho de que hasta la parte más elemental de la inducción resulte contener una parte que es informal (a saber, la división de su vocabulario en una parte proyectable y otra no-proyectable) de nuevo apoya con firmeza la conclusión sugerida por nuestra discusión del teorema de Bayes, es decir, que no puede trazarse una línea

[5] Ver su *Fact, Fiction and Forecast*, 2.ª edición, Hackett, 1977, cuya primera edición es de 1954.

divisoria entre las creencias reales de los científicos y el método científico. Goodman inventó un predicado, «verdul», que se aplica a las cosas *observadas antes del año 2000 y que son verdes o* a las cosas *que no se han observado antes del año 2000 y son azules.* Con anterioridad al año 2000 cualquier cosa examinada y verde es también examinada y verdul. Cualquier regla formal de proyección que nos diga que cuando hemos examinado cierto número de cosas, esmeraldas, por ejemplo, que tienen una propiedad *P,* podemos inferir que «Todas las esmeraldas son *P*», nos permitiría realizar las inferencias contradictorias «Todas las esmeraldas son verdes» y «Todas las esmeraldas son verdules». Y Goodman muestra convincentemente que todo intento de eliminar predicados «irrisorios» como «verdul» desde bases puramente formales no puede resolver el problema [6].

Hay una estrecha conexión, en efecto, entre la dificultad que Goodman plantea a la inducción baconiana y la necesidad de un *a priori* en relación con el teorema de Bayes. Supongamos que el científico ha de elegir (en algún momento anterior al año 2000) entre las dos hipótesis «Todas las esmeraldas son verdes» y «Todas las esmeraldas son verdules». Supongamos que la evidencia relevante reside en que se ha examinado una ingente cantidad de esmeraldas y se ha comprobado que todas son verdes (y por tanto, también se ha comprobado que todas son «verdules»). Si el científico calcula el grado de apoyo de las dos hipótesis utilizando el teorema de Bayes, entonces resulta que puede hallar o bien un grado de corroboración mucho más elevado para la hipótesis normal («Todas las esmeraldas son verdes») o bien un grado de corroboración mucho más elevado para la hipótesis anormal («Todas las esmeraldas son verdules»), o bien un grado de corroboración igual para ambas hipótesis, dependiendo del *a priori* que mantenga. Si la métrica de la probabilidad subjetiva de alguien asigna un grado mucho más elevado de probabilidad *a priori* a «Todas las esmeraldas son verdes» que a «Todas las esmeraldas son verdules», entonces se comportará de hecho como si estuviera proyectando «verde» y no proyectando «verdul». Desde un punto de vista bayesiano, la necesidad de decidir, antes de llevar a cabo una inducción, qué predicados son proyectables y qué predicados no lo son, es sólo un caso especial de la necesidad de un *a priori.*

---

[6] La solución del propio Goodman es considerar la forma *más* la historia de las anteriores proyecciones de los predicados implicados en la inferencia (junto con ciertas cuestiones relacionadas, por ejemplo, «el atrincheramiento» y «la predominancia»). Según la propuesta de Goodman se seguirá que una cultura que *siempre* ha proyectado predicados tan «irrisorios» como su famoso predicado «verdul» estaría perfectamente justificada al hacerlo —¡sus inferencias serían inductivamente válidas!—.

Aunque estoy de acuerdo con Goodman en que el ajuste con las prácticas pasadas es un importante principio de la ciencia, considero que Goodman ofrece una versión de este principio demasiado simple y demasiado relativista.

Karl Popper ha sugerido que entre hipótesis alternativas debe aceptarse *la más falsable;* pero resulta que sus medidas formales de falsabilidad producirán resultados diferentes, según los predicados del lenguaje que se escoja o se considere como primitivos. Aunque uno piense con Popper que el científico intenta encontrar *la hipótesis más falsable todavía no desechada,* o piense que el científico intenta calcular *grados de corroboración para hipótesis,* sigue siendo necesario un elemento informal correspondiente a la decisión goodmaniana de que ciertos predicados son proyectables y otros no lo son.

En este punto, el lector puede preguntarse: si no hay tal cosa como el método científico, o si el método científico, en la medida en que puede ser formalizado, depende de *inputs* que no son formalizables, entonces ¿cómo podemos dar cuenta del éxito de la ciencia? Es evidente que la ciencia ha sido una institución asombrosamente exitosa. Nos inclinamos a sentir que la razón de sus éxitos debe tener algo que ver con las diferencias en el modo en que los científicos reúnen conocimientos y el modo en que la gente tradicionalmente reunía conocimiento en las épocas precientíficas. ¿Tan erróneo es este punto de vista? La respuesta es que no. Las alternativas entre las que hemos de elegir no son que la ciencia tenga éxito porque sigue algún tipo de algoritmo formal riguroso, por una parte, y que la ciencia tenga éxito por puro azar, por otra. Comenzando por el siglo quince y alcanzando una especie de apogeo en el siglo diecisiete, los científicos y los filósofos empezaron a proponer un nuevo conjunto de máximas metodológicas. Estas máximas no son rigurosas reglas formales; su aplicación requiere racionalidad informal, es decir, inteligencia y sentido común; no obstante, configuraron y configuran el conocimiento científico. En resumen, hay un método científico, pero este presupone nociones previas de racionalidad [7]. No es un método que pueda servir como la parte más importante de la misma definición de la racionalidad.

Uno de los metodólogos más importantes del siglo diecisiete fue el físico Boyle. Con anterioridad al siglo XVII los físicos no distinguían claramente entre llevar a cabo experimentos y describir simplemente experimentos mentales que confirmarían aquellas teorías en las que creían sobre una base más o menos *a priori.* Por otra parte, los físicos no vieron la necesidad de publicar las descripciones de los experimentos fallidos. En suma, los experimentos fueron concebidos en gran parte como *ilustraciones* de doctrinas que se creían sobre bases deductivas y *a priori,* y no como evidencia contra las teorías. Boy-

---

[7] El propio Mill lo reconoce (a regañadientes) cuando escribe que no podemos esperar que el método científico funcione «si lo suponemos unido a la estupidez universal» *(El utilitarismo,* cap. 2).

le escribió manuales de procedimiento experimental, enfatizando la necesidad de ofrecer una descripción completa de todos los experimentos que se llevasen a cabo, incluyendo los experimentos fallidos. El propio Boyle fue discípulo del filósofo Francis Bacon, e indudablemente se vió inducido a apreciar la importancia de esas reglas por la postura inductivista de Bacon; no obstante, las instrucciones dadas por Boyle pueden haber tenido tanta o más importancia en la configuración del curso de la investigación científica que la defensa del procedimiento inductivo ofrecida por Bacon, más abstracta y esquemática.

Desviarse del intento de establecer teorías *a priori* hacia el intento de probar teorías mediante la derivación de conclusiones comprobables a partir de ellas y la realización de experimentos significó sin duda un cambio metodológico. Sin embargo, como ya hemos visto, no podemos identificar simplemente *ser racional* con *creer teorías sólo porque se apoyan en experimentos cuidadosamente realizados*. Por un lado, ni siquiera en la ciencia es posible siempre llevar a cabo experimentos controlados. A veces debemos confiar en la observación pasiva más bien que en el tipo de intervención activa que implica el término «experimento». Y, como vimos antes, aun cuando se hayan llevado a cabo experimentos con el propósito de elegir entre teorías alternativas, la estimación del grado en que los resultados experimentales apoyan las diversas hipótesis alternativas es, a pesar de todo, una cuestión completamente informal.

Popper ha argumentado repetidamente, contra lo que hemos estado alegando, que sí *hay* un método científico distintivo, que sí puede ser establecido y que sólo debemos confiar en éste para descubrir la naturaleza del mundo.

No obstante, Popper cree que en la toma de decisiones éticas aplicamos otros tipos de racionalidad que son más amplios que la racionalidad científica.

Según la concepción popperiana, expuesta en su influente obra *La lógica de la investigación científica* y en publicaciones subsiguientes, Popper ha argumentado que el método científico consiste en proponer teorías «sumamente falsables»: teorías que implican arriesgadas predicciones. Luego pasamos a probar todas las teorías hasta que sólo una sobrevive. Seguidamente aceptamos la hipótesis superviviente como una hipótesis con la cual proseguir hasta nueva orden, y repetimos el procedimiento entero. Y puesto que la eliminación de todas las teorías excepto una se realiza desde una base deductiva —una teoría se elimina cuando implica una predicción definitivamente falsada— no se requiere usar el teorema de Bayes, ni tampoco entra en juego ninguna estimación de grados de corrobación, afirma Popper.

El punto de vista de Popper se encuentra con el problema de que es imposible probar todas las teorías sumamente falsables. Por ejem-

plo, la teoría que sostiene que si coloco una calza de harina sobre mi cabeza y golpeo la mesa 99 veces entonces aparecerá un demonio, es sumamente falsable, pero no voy a tomarme la molestia de comprobarlo. Pero incluso aunque lo estuviera deseando, podría concebir 10^{100} teorías similares y para comprobarlas todas no sería suficiente una vida humana, ni siquiera toda la vida de la especie entera. Por tanto, por razones lógicas, es necesario seleccionar, sobre una base metodológica, un número muy pequeño de teorías que nos tomaremos la molestia de comprobar; y esto significa que hasta en el método popperiano entra en juego algo parecido a una selección previa. Como observé anteriormente, hasta los cálculos popperianos de los grados de falsabilidad son sensibles a la cuestión de cuáles son los predicados que un científico considera como primitivos en su lenguaje, y, en este sentido, hasta la noción de falsabilidad requiere una decisión previa análoga a la decisión de Goodman de que ciertos predicados son «proyectables» y otros no lo son. Renunciemos a estos puntos técnicos que en cualquier caso no interesan a nuestra presente discusión. Aun cuando el método popperiano sea incompleto, y requiera ser complementado con un método más intuitivo que no podemos formalizar en la actualidad, ¿no podría ser que describiese una condición necesaria, si no suficiente, para la racionalidad científica? ¿No podría resultar, en resumen, que una condición necesaria para la aceptabilidad de una teoría científica sea que haya sobrevivido al test popperiano? En el mismo test popperiano podría entrar en juego una selección previa de teorías a contrastar que es en sí misma informal y para la cual no tenemos un algoritmo; el cálculo de las teorías que son más sumamente falsables entraña decisiones para las que no tenemos un algoritmo; pero, a pesar de todo, podríamos insistir en que no se acepte ninguna teoría a menos que haya sido seleccionada previamente de un conjunto de teorías, todas ellas «sumamente falsables» *intuitivamente,* y a menos que todas estas teorías, exceptuando la que aceptamos, hayan sido refutadas mediante experimentos cuidadosamente realizados. En resumen, ¿no podría ser que el consejo que debemos ofrecer al científico fuese: procede como Popper te aconseja, y allí donde los métodos popperianos no puedan formalizarse, confía en tu intuición para discernir el modo en que deben interpretarse las máximas popperianas? ¿Y no podría ocurrir que el método popperiano, pese a ser en parte vago e informal, agotase no sólo la noción de racionalidad *científica* sino también de toda *racionalidad cognitiva*? Es decir, ¿no podría ser que un enunciado sea fundadamente afirmable o racionalmente aceptable si y sólo si está implicado por una teoría que pueda aceptarse sobre la base del test popperiano? La respuesta es que tal concepción de la racionalidad es demasiado estrecha *incluso para la ciencia.* En primer lugar, desestimaría la aceptación de una de las teorías mas exitosas y generalmente más admiradas: la

teoría de Darwin de la evolución por selección natural. Esta es una consecuencia que el propio Popper está dispuesto a aceptar con ecuanimidad, pero no así la comunidad científica. La teoría de la evolución por selección natural no es muy falsable, no implica unas predicciones determinadas tales que si se revelasen erróneas la teoría sería refutada. No aceptamos la teoría de la evolución porque haya sobrevivido al test popperiano, sino porque nos proporciona una *explicación plausible* de una ingente cantidad de datos, porque ha sido fértil en la sugerencia de nuevas teorías y en el enlace con los avances de la genética, de la biología molecular, etc., y porque las teorías alternativas que en efecto se han sugerido o han sido falsadas o parecen completamente implausibles por lo que se refiere al conocimiento de fondo. En resumen, aceptamos la teoría darwiniana de la evolución por selección natural como lo que Peirce llamó «abducción», o lo que recientemente se ha llamado «inferencia hacia la mejor explicación». Es precisamente este tipo de inferencia el que Popper querría *expulsar* de la ciencia, pero Popper no va a persuadir a los científicos a que abandonen teorías que no son sumamente falsables en aquellos casos en los que esas teorías proporcionan buenas explicaciones de ingentes cantidades de datos, ni en aquellos casos en los que no hay a la vista explicaciones alternativas plausibles. En realidad, como he señalado en otro sitio [8], Popper exagera la medida en que incluso las teorías de la física clásica son sumamente falsables.

Debilitamos aún más nuestra descripción del método científico al permitir la «inferencia hacia la mejor explicación» como una forma legítima de extracción de inferencias, aun cuando la «mejor explicación» inferida no sea sumamente falsable en el sentido de Popper. El método científico se ha convertido ahora en algo tremendamente vago [9]; pero, de todas formas, ya nos esperábamos algo así, en vista de los resultados formales en lógica inductiva anteriormente descritos.

¿Podría el «método científico», tan vagamente descrito, ser ahora exhaustivo? ¿Y no podría ocurrir que ningún juicio de valor sea susceptible de verificación o conformación mediante este método, ni siquiera según esta descripción tan vaga? La respuesta es que si el método científico se describe de esta simple forma: «Lleve a cabo los experimentos y observaciones tan cuidadosamente como pueda y realice entonces inferencias hacia la mejor explicación, eliminando las teorías que no pueden ser falsadas mediante experimentos cruciales», en-

---

[8] Véase «The Corroboration of Theories», en mi libro *Mathematics, Matter and Method.*

[9] Alternativamente, podemos restringir el término «método científico» para que se refiera a las aplicaciones conscientes de las máximas del procedimiento experimental, como recomendé en *Meaning and Moral Sciences,* y poner fin al intento de hacerlo tan elástico como para que englobe cualquier cosa a la que llamemos «conocimiento».

tonces es imposible discernir lo que *no es verificable* mediante un método tan vagamente descrito. Supongamos, por ejemplo, que quiero verificar el enunciado «Juan es un hombre malvado». Podría argumentar como sigue: «Se han observado los siguientes hechos: que Juan es una persona poco amable, que Juan es extremadamente egoísta, y que Juan es una persona muy cruel. Alguien poco amable, egoísta y cruel es *prima facie* una persona malvada; por tanto, Juan es un hombre malvado». Un defensor de la opinión de que los «juicios de valor» no pueden «verificarse científicamente» podría objetar dos puntos de este argumento.

Podría objetar el último paso; esto es, el paso de *Juan es cruel, Juan es poco amable, Juan es extremadamente egoísta,* a *Juan es un hombre malvado.* Es verdad que éste es un paso conceptual; la afirmación es que hay un vínculo conceptual entre ser cruel, poco amable y egoísta y ser moralmente inicuo [10]. Por supuesto, si no hay vínculos conceptuales entre los predicados morales, este paso es inválido; pero ¿por qué hemos de creer que no existan estos vínculos conceptuales?

Podría argumentarse que usar en un argumento pasos descritos como «conceptuales» no es científico; pero, sin duda, no puede mantenerse que en la ciencia no existen tales pasos. Por ejemplo, si a partir de la descripción newtoniana del sistema solar infiero el enunciado «la atracción gravitacional de la luna causa las mareas», entonces estoy empleando mi conocimiento de que hay un vínculo conceptual entre los enunciados sobre fuerzas y los enunciados de la forma *A causó B.* La palabra «causa» no aparece en la descripción newtoniana del sistema solar y de las mareas; pero yo sé que puede describirse la fuerza gravitacional que *A* ejerce sobre *B* como si fuese causada por (la masa de) *A,* en virtud de la mera comprensión de la teoría de Newton.

Por supuesto, si describimos el método científico como si consistiese en extraer «inferencias hacia la mejor explicación», o cualquier

[10] El hecho de que una verdad o una inferencia sea de la clase que denominamos «conceptual» no significa que deba ser de carácter puramente *lingüístico* (es decir, verdad en virtud de convenciones lingüísticas *arbitrarias*). Filósofos de diferentes tendencias han observado que los conceptos, los hechos y las observaciones son interdependientes. Como comentaba en el capítulo 6, los conceptos están configurados por lo que observamos o intuímos, y, al mismo tiempo, por lo que somos capaces de observar e intuir. En este aspecto, la inferencia anterior, en la que entra en juego «bueno», es exactamente análoga a la que sigue, en la que entra en juego «consciente»: «Juan está hablando inteligentemente, actuando adecuadamente y reaccionando ante lo que sucede, por lo tanto, Juan se halla consciente». Aquí el nexo conceptual consiste en que «hablando *inteligentemente*», «actuando *adecuadamente*» y «*reaccionando* ante lo que sucede» son *prima facie* razones para adscribir consciencia, del mismo modo que la poca amabilidad, el egoísmo y la crueldad son *prima facie* razones para atribuir iniquidad moral.

otra cosa, a partir de «enunciados observacionales» que están *ellos mismos en un lenguaje neutral en lo que respecta a los valores,* podemos rechazar «Juan es poco amable» y «Juan es egoísta», como enunciados de observación (aunque, en determinados casos, podría ser más fácil obtener un acuerdo con respecto a estos enunciados que con respecto a si un objeto es de *color malva,* por ejemplo). Con todo, en los escritos de los historiadores aparecen constantemente enunciados de esta índole, por ejemplo. Es muy dudoso que la historia, la psicología clínica y la descripción en lenguaje ordinario puedan evitar palabras como «amable» y «egoísta» (y sería un inmenso problema señalar dónde debe trazarse la línea: ¿tienen valor neutral «obstinado» y «furioso»? Y dicho sea de paso: ¿tiene valor neutral «retorció su brazo salvajemente»?). En cualquier caso, identificar la *racionalidad* con la racionalidad científica así descrita sería cometer una petición de principio con respecto al problema del *status* cognitivo de los juicios de valor. Sería decir que estos juicios no son racionalmente confirmables porque *son* juicios de valor, pues la racionalidad ha sido *definida* como si consistiese exclusivamente en observación pura y neutral y extracción de inferencias a partir de premisas de valor neutral. Pero ¿por qué hemos de aceptar tal definición?

9. HECHOS, VALORES Y COGNICION

En el capítulo 6 argumentaba que «todo hecho está cargado de valores y cada uno de nuestros valores lleva consigo algún hecho». El argumento consistía, en resumidas cuentas, en que *hecho* (o verdad) y *racionalidad* son nociones interdependientes. Un hecho es algo que es racional creer, o, con más precisión, la noción de hecho (o de enunciado verdadero) es una idealización de la noción de enunciado que es racional creer. Así pues, «racionalmente aceptable» y «verdad» son nociones interdependientes. Y argumentaba que ser racional implica tener un criterio de *relevancia* además de un criterio de racionalidad, y que en nuestro criterio de relevancia están supuestos todos nuestros valores. La decisión de que una imagen del mundo es verdadera (o verdadera según nuestros conocimientos actuales o «tan verdadera como la que más») y *las respuestas a las cuestiones relevantes* (así como nuestra *capacidad* para responderlas) revelan todo nuestro sistema de compromisos valorativos, sobre el cual descansan. Un ser sin valores tampoco tiene hechos.

El examen del más simple enunciado permite ver cómo nuestros criterios de relevancia implican valores, al menos indirectamente. Consideremos la oración «El gato está sobre la estera». Si alguien emitiese este juicio en un determinado contexto, emplearía recursos conceptuales —las nociones de «gato», «sobre» y «estera»— proporcionados por una determinada cultura, cuya presencia y ubicuidad revelan algo de los intereses y valores de esa cultura, y de casi todas las culturas. Tenemos la categoría «gato» porque consideramos significativa la división del mundo en *animales* y *no animales*, y además nos interesa a qué *especie* pertenece un animal dado. Es *relevante* que haya un *gato* sobre la estera, y no meramente una *cosa*. Tenemos la categoría «estera» porque consideramos significativa la división de las cosas inanimadas en *artificiales* y *no-artificiales*, y además nos interesa la *naturaleza* y el *propósito* que tiene un determinado artefacto. Es relevante que el gato esté sobre la *estera*, y no únicamente sobre una *cosa*. Y tenemos la categoría «sobre» porque nos interesan las *relaciones espaciales*.

Démonos cuenta: hemos examinado el más trivial enunciado imaginable, «El gato está sobre la estera», y nos hemos encontrado con que las presuposiciones que hacen que este enunciado sea *relevante* en ciertos contextos incluyen la significatividad de las categorías *animado/inanimado, propósito* y *espacio*. Para un espíritu que no esté dispuesto a considerar *relevantes* tales categorías «El gato está sobre

la estera» sería una observación tan *irracional* como lo sería el enunciado «El número de objetos hexagonales de esta habitación es 76», proferido a mitad de un *tete-à-tete* entre jóvenes amantes.

No sólo se muestran en nuestras categorías *(artefactos, nombres de especies, términos para las relaciones espaciales)* muchos hechos generales con respecto a nuestro sistema de valores, sino que, como veíamos en el capítulo 6, en nuestro uso de palabras clasificatorias específicas («amable», «egoísta») también se revelan nuestros valores más específicos (sensibilidad y compasión, por ejemplo). Para repetirlo una vez más, nuestros criterios de relevancia revelan todo nuestro sistema de valores, en el cual descansan.

La relevancia de esta discusión sobre la relevancia en relación con la pregunta que surgió en el capítulo precedente («¿Cuál es el valor de la racionalidad?») es inmediata. Si la «racionalidad» es una capacidad (o mejor dicho, un sistema integrado de capacidades) que permite al que la posee determinar qué *preguntas* son relevantes y qué *respuestas* justificadamente aceptables, entonces la racionalidad lleva el valor puesto. Pero no es necesario argumentar que *esta* concepción de la racionalidad está tan cargada de valor como la propia noción de relevancia.

No obstante, puede objetarse que he agrupado factores que no van juntos, encubriendo así una especie de truco de prestidigitación. Podría afirmarse que el mismo hecho de que haya hablado de dos factores, aceptabilidad racional y relevancia, testifica la persistencia de algo parecido a la dicotomía hecho-valor. Una persona racional, según la concepción que el objetor tiene en mente, sería una persona que pudiese distinguir entre lo que puede y lo que no puede afirmarse justificadamente; lo que una persona elige considerar *interesante* o *importante* o *relevante* podría resultar pertinente a la hora de evaluar su carácter o su salud mental, pero no su racionalidad *cognitiva,* según el objetor.

Sin embargo, la aceptabilidad y la relevancia son interdependientes en cualquier contexto. El uso de cualquier palabra —sea ésta «bueno» o «consciente» o «rojo» o «magnético»— supone una historia, una tradición de observación, generalización, práctica y teoría.

También nos comprometemos con la actividad de *interpretar* esa tradición, y de adaptarla a nuevos contextos, ampliándola y criticándola. Podemos interpretar la tradición de diversas formas, pero de ningún modo podemos aplicar una palabra si nos situamos totalmente fuera de la tradición a la que pertenecemos. Y el hecho de permanecer dentro de la tradición a la que pertenecemos afecta, con toda seguridad, a lo que consideramos como «aceptabilidad racional». Si existiera un método que pudiera utilizarse para verificar cualquier enunciado, sin importar qué conceptos contuviese, entonces sería factible mantener la separación propuesta entre la capacidad de verificar

un enunciado y el dominio de un conjunto *relevante* de conceptos. Pero ya hemos visto que no hay razones para aceptar el mito del Unico Método.

LA TEORIA DE LOS DOS COMPONENTES

Nuestras intuiciones actuales con respecto a la racionalidad parecen hallarse en conflicto, y ninguna filosofía parece reconciliarlas del todo. Por una parte, es sencillamente falso que *nunca* juzguemos los fines como racionales o irracionales; por otra, cuando nos enfrentamos con un caso como el del hipotético «nazi racional», no vemos cómo justificar el que desestimemos por *irracional* un sistema de fines tan inteligentemente elaborado y meditado, si bien lo encontramos moralmente repugnante.

Se ha sugerido que podríamos resolver estos problemas del siguiente modo: si asumimos una nítida dicotomía hecho-valor, podemos justificar el que tachemos de irracional al hombre que se interesa sólo en saber el número de pelos que la gente tiene en la cabeza, sobre la base de que tiene una percepción inadecuada de los *hechos* (qué significa aquí «adecuada» es un problema, desde luego). El nazi únicamente discrepa de nosotros con respecto a los *valores,* y no por ello es *irracional.* Quizá podamos tratar los casos intermedios siguiendo la línea sugerida por B. Williams. En particular, nuestro argumento contra el fetichista del método, esto es, que no podemos desechar la «evidencia observacional» de juicios descriptivos como «Juan es amable» sin caer en circularidad, puede contrastarse proponiendo la teoría que establece que el vocabulario descriptivo-moral del lenguaje ordinario tiene simultáneamente dos «componentes de significado». Un componente es fáctico: hay ciertos estándares de amabilidad generalmente aceptados, y «Juan es amable» trasmite la información de que Juan satisface esos estándares [1]. Pero también hay un componente de significado *emotivo:* «Juan es amable» transmite una «actitud favorable» hacia cierto aspecto de la conducta de Juan. Se afirma que lo que puede ser racional es la aceptación del componente fáctico del

[1] Sin embargo, es un problema saber lo que significa aquí «estándar». Si la afirmación es que «Juan es amable» es descriptivamente verdadero (es decir, el componente fáctico es verdadero) si y sólo si *la mayoría de los hablantes estarían de acuerdo en que Juan es amable,* entonces de *este* análisis del «componente fáctico» se seguiría que no podría haber una persona a quien un juez ecuánime clasificara *correctamente* como «amable», pese a que la mayoría de los hablantes *discrepase.* Tal descripción del «componente fáctico» equivaldría simplemente a la pretensión de que todas las verdades (al menos con respecto a «estándares») deben ser *públicas*; pero, si uno no es mayoritarista, ¿por qué razón ha de creerse semejante afirmación? (Como veíamos en el capítulo 5, la pretensión de que toda verdad es pública se autorrefuta.)

enunciado «Juan es amable»; la aceptación del componente de significado emotivo no es ni racional ni irracional.

La noción de «fáctico» proviene aquí de la concepción del Mobiliario del Mundo que cada filósofo prefiera. Para un filósofo materialista, el componente «fáctico» del significado de cualquier enunciado ha de consistir en un enunciado expresable en el vocabulario de la *ciencia física*. Pero de inmediato surge una dificultad parecida a la que infestó al fenomenalismo.

Recordemos que el fenomenalismo era la doctrina que sostenía que todos los enunciados con pleno significado eran traducibles a enunciados sobre sensaciones, sin que quedase residuo alguno. El «fisicalismo» (como ha llegado a llamarse el tipo de materialismo que estamos discutiendo) era la doctrina que sostenía que el significado «fáctico» completo de cualquier enunciado puede traducirse al lenguaje de la física, sin que quede residuo alguno. Y, una vez más, la doctrina parece ser falsa.

Para comprender por qué, consideremos no un enunciado descriptivo-moral, sino un enunciado psicológico, digamos «*X* está pensando en Viena». Es obvio que aun cuando haya condiciones necesarias y suficientes expresables en términos de estados cerebrales, o de lo que sea, para que un determinado individuo esté pensando en Viena, para determinar estas condiciones se necesitaría una teoría neurológica (o quizá funcionalista-psicológica) inimaginablemente perfecta. Las condiciones de verdad de este enunciado son dependientes del contexto, relativas a intereses y vagas. Así que no hay razón alguna para pensar, ni siquiera «en principio», que exista en el lenguaje de la física una expresión *finita* que sea verdadera (en cualquier mundo posible) de un *X* si y sólo si *X* está pensando en Viena. Además de que podría ser *falso* que en el lenguaje de la física exista de hecho un equivalente finito del enunciado del lenguaje ordinario «*X* está pensando en Viena», aunque tal equivalencia existiese, estaría basada en una teoría o en un grupo de teorías empíricas por ahora desconocidas (quizá sean tan complicadas que los seres humanos *jamás las conoceremos*) que sin duda no forman parte del *significado* del enunciado «*X* está pensando en Viena». En resumen, es falso que «*X* está pensando en Viena» *signifique* «*X* está en tal-y-tal estado cerebral (física o funcionalmente especificado)».

Y lo que es cierto para el enunciado «*X* está pensando en Viena», lo será para cualquier predicado del lenguaje ordinario cuyas condiciones de aplicación no se ajusten adecuadamente a las que regulan los conceptos físicos. «*X* es amable», incluso «*X* es marrón», «*X* es un terremoto» y «*X* es una persona», tampoco son *traducibles* al lenguaje de la «teoría física». Y ello significa que si hay dos componentes en el significado de «*X* es amable», la única descripción que podemos dar del «significado fáctico» de ese enunciado es que éste es ver-

dadero si y sólo si *X* es *amable*. Y ello trivializa la noción de «componente fáctico».

Decir que la teoría de los dos componentes fracasa no es negar que el enunciado «*X* es amable» tenga cierta fuerza emotiva. La tiene, mas no siempre. Como señalábamos en el capítulo 6, podemos usar el enunciado «*X* es amable» para gran variedad de propósitos: evaluar, describir, explicar, predecir, etc. La distinción entre los *usos* que pueden dársele a un enunciado no nos exige que neguemos la existencia de un *enunciado* como «*X* es amable».

MOORE Y LA «FALACIA NATURALISTA»

La actual dicotomía hecho/valor se originó en la afirmación de Weber según la cual los «juicios de valor» no pueden ser racionalmente confirmados; pero la dicotomía fue reforzada por G. E. Moore (contrariamente a sus propias intenciones). Escribiendo en un tiempo en el que Bertrand Russell y John Maynard Keynes, junto con otros futuros miembros del grupo de Bloomsbury, eran todavía jóvenes estudiantes, Moore argumentó en favor de la tesis de que el bien era una propiedad «no-natural», esto es, una propiedad que se encontraba totalmente excluida de la ontología fisicalista de la ciencia natural. Pero en su defensa del no-naturalismo le salió el tiro por la culata. Moore pudo haber convencido a sus estudiantes de que existían «propiedades no-naturales» (aunque Russell, al final, perdió la fe), pero posteriores filósofos de tendencia naturalista se inclinaron a pensar que lo que Moore había proporcionado era una *reductio ad absurdum* de la idea de que existen propiedades de valor. En los años 30, Charles Stevenson y los positivistas lógicos propusieron la «teoría emotivista de la ética», teoría en la que «*X* es bueno» significa «Yo apruebo *X*, apruébelo usted también», o algo de esta índole. Las propiedades de valor comenzaron a ser rechazadas por razones epistemológicas, y más, si cabe, por razones ontológicas; como John Mackie[2] ha expuesto recientemente, la existencia de *actitudes de valor* es compatible con la ciencia natural, pero la de *propiedades de valor* no. Estas, afirma Mackie, son «ontológicamente extrañas», es decir, son propiedades misteriosas y raras, y la gente ilustrada debería dejar de creer en su existencia.

El argumento de Moore en defensa de que *bueno* no puede ser una propiedad fisicalista (una propiedad «natural») consistía en que si «bueno» es la misma propiedad que «conducente a maximizar la uti-

---

[2] *Ethics, Inventing Right and Wrong*, Penguin, 1977.

lidad total» (o que cualquier propiedad física o funcional que usted quiera substituir), entonces

(1) «Esta acción no es buena, aunque conduce a maximizar la utilidad total»

es un enunciado *autocontradictorio,* y no sólo *falso.*

Pero ni siquiera un utilitarista pretendería que (1) es *autocontradictorio.* Y esto muestra, según Moore, que aunque ser bueno y conducir a maximizar la utilidad total podrían ser propiedades que estuviesen correlacionadas, no pueden ser la misma e *idéntica* propiedad.

No obstante, el argumento de Moore se centra en suposiciones que muchos filósofos del lenguaje rechazaríamos hoy. En primer lugar, su argumento niega implícitamente que pueda haber algo como una *identidad sintética de propiedades.* Pero, como señalé en el capítulo 4, esta imposibilidad descartaría descubrimientos científicos aceptados como el descubrimiento de que la temperatura es la misma magnitud que la *energía cinética molecular media.* (De hecho, podría utilizarse la «prueba» de Moore para mostrar que la *temperatura* debe ser una «propiedad no-natural». Pues no hay *contradicción* en «*X* tiene la temperatura *T,* pero no tiene la energía cinética molecular media *E'*» (donde *E* es el valor de la energía cinética que corresponde a la temperatura *T*) aun cuando el enunciado sea siempre falso como cuestión de hecho empírico. De modo que Moore debería de concluir que la temperatura está tan sólo *correlacionada* con la energía cinética molecular media; las dos propiedades no podrían ser literalmente *idénticas*). De hecho, Moore fundió *propiedades* y *conceptos.* Hay una noción de propiedad según la cual el hecho de que *dos conceptos* sean diferentes («temperatura» y «energía cinética molecular media», digamos) no plantea para nada el problema de que las correspondientes *propiedades* sean diferentes. (Y descubrir cuántas magnitudes físicas hay no es descubrir algo con respecto a *conceptos,* si no con respecto al *mundo.*) Puede que el *concepto* «bueno» no mantenga ninguna relación de sinonimia con ningún concepto fisicalista (después de todo, el lenguaje descriptivo-moral y el lenguaje fisicalista son versiones (en el sentido de Goodman) completamente diferentes). Pero de esto no se sigue que *ser bueno* no sea la misma propiedad que *ser P,* para alguna adecuada propiedad fisicalista (o mejor, *funcionalista*) *P.* Por lo general, un término aprendido ostensivamente para una propiedad *P* (por ejemplo «Tiene una temperatura elevada») no mantiene una relación de sinonimia con una definición teorética de esa propiedad; averiguar qué es la temperatura (y, como podría sugerir algún filósofo, averiguar qué es la *bondad*) requiere investigación empírica y teórica, y no análisis lingüístico, esto es, no requiere una reflexión en torno a los significados.

La idea de Saul Kripke de verdades «metafísicamente necesarias»

que han de ser aprendidas *empíricamente,* la idea de «verdades necesarias epistemológicamente contingentes»[3], que penetró en la filosofía del lenguaje pocos años después de que yo introdujera la «identidad sintética de propiedades», amplía e ilumina el punto en el que estoy insistiendo. La observación de Kripke, aplicada al caso de la temperatura/energía cinética, es que si alguien describe un mundo lógicamente posible en el que la gente tiene sensaciones de calor y frío, en el que hay objetos que transmiten la sensación de calor y objetos que transmiten la de frío y en el que estas sensaciones se *explican por un mecanismo distinto de la energía cinética molecular media,* entonces *no* diríamos que ha descrito un mundo posible en el cual *la temperatura no es energía molecular media.* Más bien diríamos que ha descrito un mundo en el que *algún mecanismo diferente de la temperatura hace que ciertos objetos se perciban calientes y otros fríos.* Una vez hemos aceptado el «enunciado de identidad sintética» según el cual la temperatura *es* energía cinética-molecular media (en el mundo *actual*), no hay nada que *cuente* como un mundo posible en el que la temperatura *no* es energía cinética molecular media.

Tradicionalmente se ha llamado «necesario» a un enunciado que es verdadero en cualquier mundo posible. Y «esencial» a una propiedad que algo tiene en todo el mundo posible. Saul Kripke está afirmando, sirviéndose de esta terminología tradicional, que el enunciado «La temperatura es energía cinética molecular media» *es una verdad necesaria, aun cuando no podamos saberlo a priori.* El enunciado es empírico, pero necesario. O, por decirlo con otras palabras, ser energía cinética molecular media es una propiedad esencial de la temperatura: hemos descubierto la *esencia* de la temperatura mediante investigación *empírica.* Estas ideas de Kripke han tenido gran impacto en la filosofía del lenguaje, la metafísica y la filosofía de las matemáticas; aplicadas al argumento de Moore son devastadoras. Moore argumentaba a partir del hecho de que (1) *sólo puede ser contingentemente falso,* que ser *P* (para alguna adecuada propiedad natural *P*) no podía ser una propiedad esencial de la *bondad;* y esto es lo que la nueva teoría de la necesidad impide. Todo lo que se puede inferir válidamente a partir del hecho de que (1) no es autocontradictorio es que «bueno» no es sinónimo de «conducente a la máxima utilidad» (no es sinónimo de *P,* para cualquier término *P* perteneciente a la versión fisicalista del mundo). De esta no-sinonimia de *palabras* no se sigue nada acerca de la no-identidad de *propiedades.* No se sigue nada con respecto a la esencia de la bondad.

[3] KRIPKE, *Naming and Necessity,* Harvard, 1980 (de unas conferencias dadas originalmente en Princeton, 1970).

Ruth Anna Putnam [4] ha señalado que existe un argumento común en favor de que la bondad no puede ser una propiedad natural que tampoco resuelve el problema: el argumento de que «*X* es bueno» tiene fuerza emotiva, expresa una «actitud favorable», etc.

Ruth Anna Putnam señala que este argumento no tiene en cuenta que muchos predicados descriptivos *adquieren* con naturalidad cierta fuerza emotiva. «Babeó la comida sobre su camisa» es un enunciado que tiene en nuestra cultura una fuerza marcadamente negativa, a pesar de ser literalmente una *descripción*. Cualquier palabra que represente algo que la gente de una cultura *valore* (o devalúe) tenderá a adquirir fuerza emotiva.

La palabra «bueno», en su sentido moral, se aplica a muchas cosas; algunas —por ejemplo, los buenos estados de ánimo— pueden ser valoradas por su naturaleza: puede formar parte del contenido de ese mismo estado de ánimo el que uno valore encontrarse en él. No pretendo sugerir la conversa: que cualquier estado de ánimo que se valore por su naturaleza, en este sentido, sea bueno; eso sería claramente falso. Supongamos que «bueno» se definiese de modo que las cosas que se valoran por su naturaleza, y no existe ninguna buena razón para *devaluarlas* (como la hay para devaluar ciertos estados de ánimo inducidos por las drogas, los cuales a veces se valoran por su naturaleza), cuenten como «buenas».

Entonces, es de esperar que el enunciado que afirma que algo es bueno tenga una «fuerza emocional» positiva *debido a la naturaleza de esta propiedad*. Aun cuando uno *no* sea consecuencialista (es decir, alguien que piense que *cualquier* cosa que produzca *consecuencias* suficientemente buenas, es buena), no hay duda de que la razón más común para llamar buena a una acción es que tenga buenas consecuencias, entre las cuales podría estar el que promoviese estados o situaciones que se valoran por su naturaleza; una vez más, la misma naturaleza de la propiedad explica por qué la descripción llega a tener una fuerza emotiva «favorable».

Mackie defiende que la bondad es una propiedad ontológicamente «extraña» introduciendo como premisa la suposición de que no es posible *saber* que algo es bueno *sin* tener una actitud «favorable» hacia ese algo. Y ello equivale a suponer el *emotivismo* para probar el emotivismo. Con frecuencia se representa a los diablos del averno usando la palabra «bueno» con una fuerza emotiva *negativa* («Tiene una deplorable tendencia a la bondad moral», podría decir uno de ellos); al contrario que Mackie, no encuentro que estos usos sean

[4] «Remarks on Wittgenstein's Lecture on Ethics», en Haller *et al* (eds.), *Language, Logic and Philosophy, Proceedings of the 4th Intern. Wittgenstein Symposium*, Viena, 1980.

lingüísticamente impropios, o que impliquen alguna contradicción. ¿Acaso nunca hemos oído decir: «Sé que está mal hacerlo, pero no me importa»? Como ha señalado Philippa Foot, uno puede rechazar los llamamientos a la moralidad diciendo «Yo no aspiro a ser un hombre bueno». La mera posibilidad de estas proferencias muestra que aunque existe en realidad una diferencia entre *el uso descriptivo* del lenguaje y el uso *prescriptivo o aprobativo,* tal diferencia no está en función del *vocabulario.* Las palabras «descriptivas» pueden usarse para alabar o denostar («Babeó la comida sobre su corbata») y las palabras «evaluativas» pueden usarse para describir y explicar. (Consideremos el siguiente diálogo: «Juan debe ser un hombre excepcionalmente bueno para hacer lo que hizo». «No, él nunca ha sido un dechado de moral, sino más bien lo contrario; pero debe tener una capacidad de autosacrificio que jamás sospechamos que tuviera».) El lenguaje *moral* se usa en este punto con una función *explicativa.* Repitiendo una vez más lo que señalaba Ruth Anna Putnam, del hecho de que «es bueno» se utilice para aprobar no se sigue que la bondad no sea una propiedad.

La profesora Putnam señala que el argumento de Mackie acierta en una cosa, sin embargo. Es innegable que algunas expresiones morales llevan incorporada cierta orientación hacia la acción. «Yo debería», «correcto» y «deber» son los principales ejemplos de expresiones guía-para-la-acción. Como ella señala, el problema del «es/debe» no coincide con el problema del «hecho/valor». «No tengo la intención de hacer lo que debería hacer» nos suena más extraño que «No tengo la intención de ser un hombre bueno» (y «No tengo la intención de hacer lo que debo» parece un disparate)*.

Mackie señala que ninguna propiedad física tiene incorporada una conexión con la acción (o con la aprobación de la acción) y concluye que «ser lo correcto a realizar», etc., son «ontológicamente extrañas». Pero su argumento, además de depender de la suposición de que la versión fisicalista del mundo es la teoría verdadera, prueba demasiado. Pues algunos predicados *epistémicos* (por ejemplo «racionalmente aceptable», «creencia justificada») son también guía-para-la-acción (considerando «acción» en el sentido amplio, de modo que aceptar un enunciado cuenta como un tipo de acción). Alguien puede decir «Está bien hacer *X*» y «Hay bastante evidencia de que *Y*» y no comprometerse a hacer o prescribir *X* o a aceptar *Y;* pero si alguien dice «En esta situación es correcto realizar la acción *X*» o «Creo que *X*

---

* Es difícil reflejar en el vocabulario español el carácter anómalo que el autor atribuye a este último enunciado («I am not out to do what I must»). «I must» es una expresión cuyo uso tópico conlleva el compromiso del hablante con una obligación plena, voluntaria y autónomamente asumida. Véase R. HUDDLESTON, *Introduction to the grammar of English,* Cambridge University Press, 1984, pp. 167-168 (*N. del T.*).

está completamente justificado», entonces se orienta a hacer *X* (o prescribir *X*) y a aceptar *Y*. «Justificado» (en el caso de las creencias) tienen la característica de ser guía-para-la-acción, como la tiene «correcto», en sentido moral.

Si remedamos ahora el argumento de Mackie y concluimos que no existe tal propiedad como *estar justificado,* sino sólo «actitudes de justificación», entonces nos estrellamos en un total relativismo. Antes de ir tan catastróficamente lejos, podemos detenernos a considerar por qué los predicados guía-para-la-acción son «ontológicamente extraños» para un fisicalista comprometido.

En el capítulo 2 argumentaba que a semejante fisicalista ha de parecerle ontológicamente extraña hasta la propia *referencia.* Si todo lo que realmente existe son propiedades y relaciones fisicalistas, entonces la referencia, para existir, debe ser una relación fisicalista; pero entonces el problema es la sobreabundancia de candidatos. Hay un número infinito de relaciones de referencia admisibles (y todas ellas son fisicalistas, o al menos naturalistas, si admitimos la teoría de conjuntos como parte de una versión naturalista del mundo). Si la relación de referencia fuera una de éstas, entonces ese hecho sería en sí mismo un hecho metafísico último de muy extraña índole.

Lo que convierte a este hecho en algo tan extraño es que hemos incorporado cierta *neutralidad,* cierta inconsciencia o carencia de deliberación *(mindlessness)* en nuestra misma concepción de la naturaleza. Se supone que la naturaleza no tiene intenciones, intereses o puntos de vista. Dando esto por cierto, ¿cómo podría estar seleccionada metafísicamente una relación de referencia admisible?

Esa misma inconsciencia de la naturaleza hace parecer «extraños» los predicados guía-para-la-acción «Es correcto» y «Es una creencia justificada». Sería extraño que una propiedad fisicalista *P* fuera idéntica a la corrección moral o a la justificación epistemológica —precisamente por la misma razón por la que sería «extraño» que la referencia fuera una relación fisicalista. Sería como si la naturaleza tuviese valores, en el caso moral, o intenciones referenciales, en el caso semántico.

Por esta razón, pienso que Moore estaba en lo cierto (aun cuando sus argumentos no son aceptables) al sostener que «bueno», «correcto» (y también «creencia justificada», «refiere» y «verdadero») no son idénticas a propiedades y relaciones fisicalistas. Y lo que *esto* muestra no es que no existan la bondad, la corrección, la justificación epistemológica y la verdad, sino que el naturalismo monista (o «fisicalismo») es una filosofía inadecuada.

EL «NAZI RACIONAL», OTRA VEZ

Lo que nos preocupaba anteriormente era que no veíamos cómo argumentar que el hipotético nazi «perfectamente racional» tuviese fines irracionales. Quizá el problema sea el siguiente: identificamos con demasiada simpleza el problema de la racionalidad del *nazi* (como alguien que posee una cosmovisión o cosmovisiones) con la racionalidad de sus *fines*. Si no hay ningún fin «en» el nazi al que podamos apelar, entonces parece extraño diagnosticar la situación diciendo «Karl tiene metas irracionales». Aunque esto sea parte de lo que finalmente concluyamos, seguramente lo *primero* que queremos decir es que Karl tiene metas *monstruosas,* no que tiene metas irracionales.

Pero el asunto a considerar, si es que vamos a discutir la racionalidad de Karl, es la irracionalidad de sus *creencias y argumentos,* no la de sus fines.

Supongamos primero que Karl pretende que las metas nazis son moralmente correctas y buenas (como los nazis pretendieron de hecho, si bien no en los ejemplos de los filósofos). Entonces, *de hecho,* estará diciendo tonterías. Asentirá a todo tipo de proposiciones «fácticas» falsas, por ejemplo, que las democracias están bajo control de una «conspiración sionista», y emitirá proposiciones morales (por ejemplo, que un «ario» tiene la obligación de subyugar las razas no arias a la «raza superior») para las que no tienen ningún buen argumento. El concepto de «buen argumento» al que estoy apelando pertenece al discurso moral ordinario; pero es la noción adecuada, en caso de que el nazi trate de justificarse *dentro* del discurso moral ordinario.

Supongamos, por otra parte, que el nazi repudia en su conjunto todas las nociones morales ordinarias (como hacía nuestro hipotético superbenthamita). Como argumentaba anteriormente, una cultura que repudiase las nociones morales ordinarias, o las sustituyese por nociones derivadas de una ideología y de una perspectiva moral diferente, perdería la capacidad de describir adecuada y perspícuamente las relaciones interpersonales cotidianas, los acontecimiento sociales y los acontecimientos políticos *según el estado actual de nuestros conocimientos.* Cuando esa diferente perspectiva moral e ideológica es *superior* a nuestro actual sistema moral, esta sustitución puede ser saludable y juiciosa, por supuesto. Pero si son *nocivas,* en particular, si son pervertidas y monstruosas, entonces el resultado será simplemente una inadecuada, poco perspicaz y repulsiva representación de los hechos interpersonales y sociales. «Inadecuada», «poco perspicaz» y «repulsiva» reflejan juicios de valor, desde luego; pero ya he argumentado que la elección de un esquema conceptual refleja *necesariamente*

juicios de valor; y la racionalidad *cognitiva* es el factor decisivo en la elección de un esquema conceptual.

Aun cuando el individuo nazi no haya perdido la capacidad de utilizar nuestro actual vocabulario descriptivo-moral, aun cuando retenga en algún lugar de su cabeza las viejas nociones (como quizá algunos eruditos todavía estén familiarizados con la noción medieval de «caballería», y sean capaces de utilizarla) a pesar de todo, no empleará estas nociones (nuestras nociones descriptivo-morales, «amable», «compasivo», «justo», e «imparcial») al vivir su vida: no figurarán realmente en su construcción del mundo.

Deseo recalcar una vez más que no afirmo que lo *malo* de ser nazi sea que puede llevar a alguien a tener *creencias* perversas e irracionales. Lo malo de ser nazi son las acciones a las que ello nos conduce. El nazi es maligno y además tiene una cosmovisión irracional. Estos dos hechos están interrelacionados; pero eso no significa que el nazi sea maligno principalmente *porque* tiene una cosmovisión irracional, en el sentido de que la irracionalidad de su cosmovisión *constituya* esa *malignidad.* Pese a todo, me parece que hay un sentido en el que podemos hablar aquí de metas racionales e irracionales: tenemos derecho a llamar irracionales a aquellas metas que si se aceptan y persiguen, conducen o bien a ofrecer en su favor argumentos disparatados o irracionales (si se acepta la tarea de justificarlos dentro de nuestro esquema conceptual normal) o bien aceptar un esquema alternativo para representar hechos descriptivo-morales ordinarios (por ejemplo, que alguien es compasivo) que es irracional. Después de todo, existe una relación entre emplear un esquema conceptual racional en la descripción y comprensión de hechos moralmente relevantes y tener ciertos tipos generales de metas, opuestos a otros.

«¿Pero qué ocurre si el nazi no da razón alguna de por qué es nazi excepto «hago lo que me da la gana»? Esta es una pregunta natural, pero sin duda la respuesta natural también es la correcta: en tal caso, la conducta del nazi, además de ser atroz, sería también conpletamente *arbitraria.* Démonos cuenta que «arbitraria» es una de las palabras a las que hemos llamado descriptivo-morales, es decir, una palabra que puede usarse, sin cambio en su denotación, para evaluar (en este caso, para condenar), para describir («Juan decidió de forma completamente arbitraria cambiar de empleo»), para explicar (o para indicar que no puede ofrecerse una explicación de cierto tipo), etc. En realidad, cuando decía que la decisión de Karl de convertirse en nazi (en el caso descrito) sería completamente arbitraria, estaba principalmente *describiendo,* no evaluando. Muchas cosas de las que hago son arbitrarias, en un sentido completamente literal —por ejemplo, elegir una senda a través del campus y no otra—, pero ello no significa que en esas acciones haya algo erróneo. (Sencillamente esos asuntos son demasiado triviales.) Aun cuando haga algo importante

de forma «arbitraria» —cambiar de empleo, por ejemplo—, si no tengo responsabilidades familiares, etc., estoy en mi pleno *derecho*. Pero si la acción *requiere justificación*, entonces si alguien la lleva a cabo de modo arbitrario, sin ofrecer ninguna justificación, se expondrá a una legítima condena. Tomar una decisión que afecta adversamente y en gran medida a las vidas de los demás (y quizá a la vida de uno mismo) sin ninguna justificación, sólo como un acto arbitrario y testarudo (¡otro par más de palabras descriptivo-morales!), es un ejemplo paradigmático de irracionalidad; y no sólo de irracionalidad, sino de terquedad.

En el capítulo 7 iniciamos nuestra discusión examinando la afirmación de Bentham de que, «prejuicios aparte», el juego de las clavijas (un antiguo juego infantil similar al juego de la pulga) es exactamente tan bueno como «las artes y las ciencias de la música y la poesía». En opinión de Bentham, la única razón por la que la poesía es mejor que el juego de las clavijas es, en última instancia, el puro hecho de que la poesía ofrece mayor satisfacción que este juego (u ofrece satisfacción a un mayor número de gente, o ambas cosas). Esta opinión se equivoca básicamente en dos cosas: la primera es que «la satisfacción» (o el «autointerés») no puede ser por sí misma un objetivo de un ser que no tenga además *otros* objetivos. Si no tuviese *otro* objetivo que mi «bienestar», mi «bienestar» sería una noción asignificativa —una cuestión que se remonta al obispo Butler. Y lo que es más importante, algunas *satisfacciones* son mejores y «más nobles» que otras, y se pueden dar razones. La poesía y la música nos ofrecen solaz, amplían nuestra sensibilidad, proporcionan importantes modos de autoexpresión a mucha gente, incluyendo a la mayoría de personas de talento que ha producido la raza humana.

Que se califique de «prejuicios» a estas razones para dar mayor valor a ciertas satisfacciones frente a otras es algo que se encuentra estrechamente relacionado con la teoría de los dos «componentes», así como con la idea de que las propiedades de valor son «ontológicamente extrañas». Bentham opera con el modelo de los «hechos neutrales» y los «prejuicios» arbitrarios. En realidad, llamar «prejuicio» a la preferencia por la poesía es sólo la forma que tiene Bentham de sugerir que la *única* razón que no es «arbitraria» en la comparación de la poesía con el juego de las clavijas es el hecho de que la poesía ofrece mayor satisfacción; cualquier preferencia de un *tipo* de satisfacción sobre otro es arbitraria (esto es lo que sugiere). Pero esto es sencillamente falso, dado el lugar que la noción de preferencia «arbitraria» ocupa realmente en nuestro esquema conceptual; y si «arbitrario» se arranca del esquema conceptual al que pertenece, la afirmación carece de significado. (De modo parecido, el enunciado «Preferir la poesía al juego de las clavijas es un prejuicio» es *literalmente falso*.) Se ha sugerido que es ontológicamente legítimo admitir que

existan cosas tales como *satisfacciones,* pero que no lo es admitir cosas tales como la ampliación de la sensibilidad o de los repertorios de significados y metáforas, modos de expresión y de autorrealización, etc. La idea de que los valores no son parte del Mobiliario del Mundo y la idea de que los «juicios de valor» son expresiones del «prejuicio» son las dos caras de la misma moneda.

Hemos investigado la cuestión de si los «juicios de valor» pueden apoyarse racionalmente. Hemos visto que las diversos respuestas negativas descansan en suposiciones filosóficas dudosas: que la misma racionalidad sólo es buena para la «predicción» o para obtener «consenso», o que sólo hay un método para llegar a la verdad (a veces se afirma que el único criterio de verdad lo constituyen la predicción y el consenso), o que los juicios de valor tienen «dos componentes de significado» o que las propiedades de valor son «ontológicamente extrañas». La posición que he defendido es que cualquier elección de esquema conceptual presupone valores, y la elección de un esquema conceptual para describir las relaciones interpersonales y los hechos sociales, por no mencionar la reflexión sobre el plan de vida de uno mismo, implica, entre otras cosas, los valores *morales* que uno mantiene. No puede elegirse un esquema que simplemente «copie» los hechos, ya que *ningún* esquema conceptual es una mera «copia» del mundo. El contenido de la misma noción de verdad depende de los criterios de aceptabilidad racional, y éstos, a su vez, presuponen nuestros valores, sobre los que descansan. Expresándolo esquemática y brevemente: la teoría de la verdad presupone la teoría de la racionalidad, que a su vez presupone nuestra teoría de lo bueno.

La «teoría de lo bueno», sin embargo, no es sólo programática, sino que depende de suposiciones acerca de la naturaleza humana, de la sociedad y del universo (incluyendo suposiciones teológicas y metafísicas). Hemos tenido que revisar nuestra teoría de lo bueno (tal como está) una y otra vez, según va aumentando nuestro conocimiento y va cambiando nuestra cosmovisión.

Creo que es evidente que no existe algo como un «fundamento» para la concepción que defiendo. Y en este punto la gente se inquieta: ¿no nos hallamos muy próximos al punto de vista de que no hay diferencia entre «justificado» y «justificado para *mis* conocimientos» (una especie de solipsismo)?

En realidad, nos vemos abocados a la posición del solipsista si intentamos permanecer fuera del esquema conceptual al que pertenece el concepto de racionalidad y pretendemos ofrecer simultáneamente una noción maś «racional» de racionalidad. (Muchos pensadores han caído en el error de Nietzsche: decirnos que tenían una moralidad «mejor» que la de toda la tradición; pero lo único que produjeron, en cada caso, fue una monstruosidad, pues todo lo que *podían* hacer era

arrancar algunos valores de su contexto, *arbitrariamente,* mientras ignoraban los demás.)

Solamente podemos tener la esperanza de producir una *concepción* más racional de la racionalidad o una *concepción mejor de la moralidad si operamos dentro* de nuestra tradición (con los ecos del ágora griega, de Newton, etc., en el caso de la racionalidad, y con los ecos de las escrituras, de los filósofos, de las revoluciones democráticas, etc., en el caso de la moralidad); pero ello de ningún modo significa que en nuestras concepciones actuales todo esté bien y todo sea razonable. No estamos atrapados en infiernos solipsistas individuales, sino invitados a tomar parte en un diálogo genuinamente humano, un diálogo que combine la colectividad con la responsabilidad individual.

¿Tiene este diálogo un término ideal? ¿Hay una concepción *verdadera* de la racionalidad, una moralidad *verdadera,* aun cuando todo lo que tenemos son nuestras *concepciones* de éstas?

En este punto, la opiniones de los filósofos, como las de los demás, se dividen. Richard Rorty, en su discurso presidencial[5] a la *American Philosophical Association,* optó con firmeza por el punto de vista de que sólo existe el diálogo; no puede postularse ningún fin ideal, ni tampoco sería necesario. Pero la afirmación de que «solo existe el diálogo» ¿difiere en algo del relativismo que se autorrefuta, discutido en el capítulo 5? El mismo hecho de que hablemos de nuestras diferentes concepciones como diferentes concepciones de la *racionalidad* postula un *Grenz-begriff,* un concepto límite de verdad ideal.

---

[5] «Pragmatism, Relativism and Irrationalism», *Proceedings and Adresses of the American Philosophical Association,* August 1980. Véase también *Philosophy and the Mirror of Nature,* Princeton University Press, 1979.

APENDICE

He aquí el teorema al que hice referencia en el capítulo 2.

TEOREMA: Sea L un lenguaje con los predicados F_1, F_2, ... , F_k (no necesariamente monádicos). Sea I una interpretación que asigna una intensión a cada predicado de L. Entonces, si I es no-trivial, en el sentido de que al menos un predicado tiene una extensión que ni es vacía ni universal al menos en un mundo posible, existe una segunda interpretación J que no coincide con I, pero que satisface las mismas oraciones que I en cada mundo posible.

PRUEBA: Sean W_1, W_2, ... todos los mundos posibles, en alguna ordenación adecuada, y sea U_i el conjunto de todos los individuos posibles que existen en el mundo W_i. Sea R_{ij} el conjunto que constituye la extensión del predicado F_i en el mundo posible W_j de acuerdo con I (si F_{ij} es no-monádico, entonces R_{ij} será un conjunto de n_i-tuplos, donde n_i es el número de lugares de argumento de F_i). La estructura $[U_j; R_{ij} (i = 1,2, ..., k)]$ es el «modelo proyectado» de L relativo a I en el mundo W_j, esto es, U_j es el universo de discurso de L en el mundo W_j y (para $i = 1,2, ..., k$) R_{ij} es la extensión del predicado F_i en W_j.

Si al menos un predicado, por ejemplo F_u, tiene una extensión R_{uj} que ni es vacía ni agota U_j, seleccionemos una permutación P_j de U_j tal que $P_j(R_{uj}) \neq R_{uj}$. De lo contrario, sea P_j la identidad. Ya que P_j es una permutación, la estructura $[U_j; P_j(R_{ij}) (i = 1,2,..., k)]$ es isomorfa a $[U_j; R_{ij} (i = 1,2,..., k)]$ y, de este modo, es un modelo para las mismas oraciones de L (es decir, para las oraciones de L que son verdaderas bajo I en W_j).

Sea J la interpretación de L que asigna al predicado F_i $(i = 1,2, ..., k)$ la siguiente extensión: la función $f_i (W)$, cuyo valor en cualquier mundo posible W_j es $P_j(R_{ij})$. En otras palabras, la extensión de F_i en cada W_j bajo la interpretación J se define como $P_j(R_{ij})$. Ya que $[U_j; P_j(R_{ij}) (i = 1,2,...,k)]$ es un modelo para el mismo conjunto de oraciones que $[U_j; R_{ij} (i = 1,2,..., k)]$ (por el isomorfismo), en cada mundo posible son verdaderas las mismas oraciones bajo I y bajo J, y J difiere de I en cada mundo posible en el que al menos un predicado tiene una extensión no trivial, q.e.d.

COMENTARIO: Si, en un mundo W_j dado, hubiera dos conjuntos disjuntos que fueran las extensiones de los predicados de L en W

bajo I —por ejemplo, el conjunto de los gatos y el conjunto de los perros—, entonces, si hubiera más perros que gatos (respectivamente, más gatos que perros) podemos tomar algún conjunto de perros que tenga la misma cardinalidad que el de los gatos (respectivamente, algún conjunto de gatos que tenga la misma cardinalidad que el de los perros) y elegir un P_j que proyecte el conjunto de perros seleccionado sobre el de gatos (respectivamente, el conjunto seleccionado de gatos sobre el de perros) y viceversa. Ello nos asegurará que la extensión del primer predicado bajo J —aquél cuya extensión bajo I es el conjunto de los gatos— es en W_j el conjunto de los perros; o la extensión del segundo predicado bajo J —aquél cuya extensión bajo I es el conjunto de los perros— es en W_j el conjunto de los gatos.

SEGUNDO COMENTARIO: Si alguien desea que *no* se permute algún tipo de objetos, las «sensaciones», por ejemplo —puesto que considera que los predicados de esos objetos son «absolutos», en algún sentido—, ha de estipular simplemente que la permutación de P_j sobre esos objetos sea la identidad. Esto hará que la restricción de cualquier predicado de L para estos objetos privilegiados sea la misma bajo I y bajo J en cada mundo posible.

TERCER COMENTARIO: Ya que las oraciones reciben unas condiciones de verdad lógicamente equivalentes bajo I y bajo J, se sigue que según la «semántica de los mundos posibles» estándar, también se preservan los *condicionales contrafácticos.*

INDICE DE NOMBRES Y CONCEPTOS

La estrategia con la que Hilary Putnam se enfrentará en esta obra a algunos de los problemas más perseverantes de la filosofía —la naturaleza de la verdad, del conocimiento y de la racionalidad— se desprenderá de un diagnóstico previo: parte del descrédito del análisis filosófico se debe al estancamiento del debate entre facciones opuestas, caracterizadas por cierto número de dicotomías tradicionalmente presentadas como cánones a la hora de establecer formulaciones válidas y de determinar soluciones permisibles de aquellos problemas. El objetivo del autor será disolver estas presuntas escisiones, o al menos debilitar su fuerza coercitiva, mostrando las limitaciones internas de ambos polos y confiando en delimitar de este modo un ámbito en el que construir lo que él llama «una descripción equilibrada y humana del alcance de la razón». Putnam intenta articular la interdependencia entre las nociones de *verdad* y de *racionalidad* en una posición denominada «realismo interno», enfrentada tanto a la ingenuidad de las teorías de la verdad-copia («El punto de vista del Ojo de Dios») como a la indolencia del relativismo epistemológico

Gran parte del atractivo de la obra de Hilary Putnam descansa en su capacidad de dialogar con las tradiciones más diversas. Con todo, es de destacar el reconocido influjo que sobre su pensamiento han ejercido autores como W. V. Quine y N. Goodman.

Filosofía y Ensayo

tecnos